

Will the regulator please stand up

It's time for the South Korean government to launch an investigation into how eggs were obtained for a ground-breaking stem-cell experiment.

Once again, Seoul National University's Woo Suk Hwang is this week being accused of possible impropriety in allegedly obtaining human eggs for the first experiment to derive human stem cells from a cloned human embryo.

His accuser this time is Gerald Schatten of the University of Pittsburgh, a long-time collaborator of Hwang's. In a statement on 12 November announcing that the collaboration will now end, Schatten cites charges, first aired in *Nature* in May last year, of "oocyte donation irregularities" at Hwang's laboratory (see *Nature* 429, 3; 2004).

There were calls for an investigation back then, but South Korea's handful of bioethicists had no leverage, and nothing happened (see *Nature* 429, 490; 2004). Much of the Korean media repeated and endorsed Hwang's denials. Far from launching an investigation, the government backed his research with generous funding and dedicated a postage stamp to him. Some politicians even pledged to spearhead a drive to win him a Nobel prize.

Stem-cell researchers worldwide were scarcely more critical, perhaps fearing that any suggestion that this high-profile research had rested on an unethical practice would stain a field that has enough controversy attached to it already. As the situation in Japan amply demonstrates, such fears can rapidly thwart research opportunities in this sphere (see page 262 of this issue).

Schatten's actions reopen the questions raised last year. Did the experiment use eggs donated by a graduate student or by a member of the research team? Did donors receive payment for their eggs? Hwang has vigorously denied these allegations.

But this time, it will be harder for the Korean authorities to ignore these questions. The Korean media is taking a more critical view. According to some reports, Ky Young Park, the president's adviser for science and technology, has already promised an investigation.

An investigation led by Park would be less than optimal, however, as she was a co-author on the Hwang paper (*Science* 303, 1669–1674; 2004). She subsequently described her role in the work as that of a 'bioethics consultant' — and told *Nature* that she hadn't given any thought to the ethics of egg donation.

Park's real role in the work remains something of a mystery. Almost anyone else would be better placed to investigate this episode, but it remains to be seen who will do it. The ministry of science and technology does not seem to be keen. As time passes, an inquiry may become more difficult to conduct.

A thorough investigation is nonetheless required, not just for the sake of scientific integrity in South Korea, but to help persuade sceptics worldwide that research on human embryonic stem cells is being done ethically. This field of research could yet prove to be immensely fruitful, but it requires strong public support.

Stem-cell researchers will now find themselves on the defensive in proving that they are ready to stick to strict ethical codes. Just when Hwang was tying together an international stem-cell network with his laboratory at its hub (*Nature* 437, 1077; 2005), these allegations will reverberate around the world of developmental biology.

To maintain public support for any controversial field of science, researchers need to follow strict ethical guidelines — and be seen to be doing so. If for whatever reason that doesn't happen, responsibility jumps up a level. It then becomes the job of regulatory bodies and funding agencies to ensure that researchers are brought to account. Is anyone in South Korea going to step up to the task? ■

"To maintain public support, researchers need to follow strict ethical guidelines — and be seen to be doing so."

Heavy weather

Washington DC still doesn't seem to understand the threat posed by global warming.

Climate change is a political science, and a messy one at that. This issue of *Nature* includes overviews and opinions that shed light on how researchers and citizens are responding to the regional effects of climate change. But at any level, the field is beset with genuine scientific uncertainties and complexities. Politically, these challenges are compounded by confusion on the part of the public and manipulation by sceptics of global warming.

The United States, of course, is rife with both confused citizens and vocal sceptics. But it is also home to many of the world's leading climate scientists, and they are involved in a major attempt to take

the lead in this arena — an effort that now seems, unfortunately, to be foundering.

This week the US Climate Change Science Program held a workshop to assess its progress so far, and to look ahead to its future goals. The programme is supposed to produce 21 reports summarizing various aspects of climate science (see *Nature* 436, 890; 2005). These should represent the best consensus that science can offer, and are due to be signed off by the US government, with the White House being the final step in the approval process. But if the brief history of the project's first report is any guide, the exercise will be lucky if it ever reaches fruition.

Climate researchers had hoped that this week's meeting would showcase the science coming out of the first report, on temperature trends in the troposphere. The report had successfully undergone review by the National Research Council and was about to be posted on the web for 45 days of public review. After that, with any changes



duly incorporated, it would be sent to the White House for approval.

But this autumn, almost two years after they began work on the report, the authors were informed of a fresh requirement — that they be approved as governmental advisers under the terms of the Federal Advisory Committee Act. In theory, this extra layer of bureaucracy is meant to ensure the legitimacy of those who act as advisers. In practice, it meant that the climate scientists were fingerprinted and had their financial backgrounds checked. During this process, the report's authors were not supposed to speak to each other for several months, while their report languished.

Meanwhile, internal bickering broke out into the open. Group member Roger Pielke, a climatologist at Colorado State University, withdrew from the panel, claiming that his views that land-use changes contributed substantially to climate change were being suppressed (see *Nature* 437, 9; 2005).

Even under such conditions, science will out. Three papers based on the tropospheric temperature report have already been published in *Science* (doi:10.1126/science.1114772; doi:10.1126/science.1114867; doi:10.1126/science.1115640; 2005). The researchers are now obtaining clearance to act as governmental advisers. And on 16 November, the three lead authors of the *Science* papers were due to discuss the

findings at a seminar being held for congressional staff by the American Meteorological Society in Washington DC.

One of the researchers, climate modeller Benjamin Santer of Lawrence Livermore National Laboratory, had not been on Capitol Hill for a decade. In 1995 he was subjected to severe and unjustified criticism for his participation in that year's report from the Inter-governmental Panel on Climate Change — its first report to state that humans were having a discernible effect on the climate. Santer became the target at which climate sceptics took aim.

Santer's willingness to return to the fray is commendable. Global-warming sceptics still hold far too much sway in Washington, where one congressman earlier this year summoned novelist Michael Crichton to testify as a 'scientific' witness on climate change because of his pseudoscientific novel *State of Fear*.

In the face of such attitudes, researchers must stay the course. The government needs to streamline and accelerate the flow of information through the climate-change programme. The Bush administration owes the US public that much at least. ■

"One congressman summoned novelist Michael Crichton to testify as a 'scientific' witness on climate change."

Pulling together

Protests by Chinese students at Yale University are highlighting strains on a symbiotic relationship.

Chinese students and scientists are playing an increasingly important role in US laboratories. According to the New York-based Institute of International Education, US academic institutions are now home to some 80,000 Chinese nationals, many of them in the sciences.

They are attracted to the United States, in the main, because of its excellent research universities, which are delighted to recruit well-trained and hard-working Chinese nationals. But as our News Feature on page 278 demonstrates, reality doesn't always quite meet the visitors' expectations.

Xuemei Han, a Chinese national, was admitted by Yale University to study ecology on the basis of her strong academic background. But the language barrier, funding problems and bureaucratic tussles ultimately led to a public falling-out, which quickly escalated to become a focal point for major protests by a large number of Chinese students at Yale.

Thanks to Yale's academic reputation and its unusually well-organized graduate student body, this chain of events has won widespread attention. But it is hardly unprecedented, and enquiries by *Nature* reaffirm that many Chinese researchers feel out of step with their supervisors or their institutions.

Some supervisors may be tempted to dismiss these complaints as the usual belly-aching from the lower echelons of the laboratory. After all, it is certainly true that many of the problems encountered by Chinese graduate students are shared by their colleagues, both US-born and foreign.

But there are some issues that are particularly acute to Chinese

graduate students — by far the largest such immigrant group in the United States. One of these is the language barrier, which can be formidable for students who have often received years of written language training at home but may speak English haltingly at first. Some students also come from an academic environment where dissent is rare, and may fail to assert themselves as readily as their US colleagues. Many Chinese students interviewed by *Nature*, for example, were reluctant to give an opinion, even privately, about how their laboratories ought to be run. Finally, Chinese students have faced strenuous visa restrictions that can complicate their travel arrangements and engender insecurity about their status in the United States. Taken together, these factors can leave them feeling more isolated and disaffected than their US-born counterparts.

The principal investigators who are directly responsible for supervising the students should be aware of these concerns, and, where necessary, take appropriate actions to address them. They should make sure that students have the resources available to improve their language skills. Obviously, they should encourage everybody in the laboratory to bring forward their own ideas. And they should be patient with people who face the logistical challenges of visas and international travel back home.

Today's scientific workforce is highly mobile, and while many Chinese students and scientists will no doubt complete outstanding careers in the United States, many others will choose to return home and build up their own laboratories there. In the decades to come, these laboratories will become globally important. The experience of Chinese students and young scientists in the United States will set the tone for scientific relations between the world's only superpower and its emerging rival. ■

"Issues particularly acute to Chinese students leave them feeling more isolated and disaffected than their US-born counterparts."

RESEARCH HIGHLIGHTS

Hard wired

Cell **123**, 477–491 (2005)

US neuroscientists have cracked part of the molecular code that controls the development of nerve cells in the spine.

The researchers, led by Thomas Jessell at the Howard Hughes Medical Institute, New York, studied how motor neurons in chick embryos connect to muscles in the developing wing. They found that the expression of different combinations of the chick's 39 *Hox* genes — which determine where limbs and other body parts grow, among other things — told developing cells what kind of neuron to become and which wing muscles to connect to.

The team suggests that other kinds of spinal nerves may also use the *Hox* code, and that deciphering it completely might help them to understand the complex wiring that the spinal cord uses to control movement.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

CHEMICAL BIOLOGY

Seeing the light

Science **310**, 1006–1009 (2005)

The molecular event that initiates sight happens on a timescale that makes the blink of an eye seem an eternity. But a technique that uses laser pulses that last in the order of 10^{-15} seconds has revealed how the visual pigment rhodopsin captures the energy of incoming light.

Richard Mathies of the University of California, Berkeley, and his colleagues used femtosecond-stimulated Raman spectroscopy to observe how retinal — the light-absorbing component of rhodopsin — transforms into its structural isomer. This is the only light-sensitive step in vision.

They find that the 'wagging' of hydrogen

atoms that are attached to retinal's backbone rapidly converts the system into a new electronic ground state. The changes in the charge distribution then drive the slower structural rearrangement.

ANIMAL BEHAVIOUR

Undercover learning

Curr. Biol. **15**, 1931–1935 (2005)

Wood crickets (*Nemobius sylvestris*; pictured left) have surprised biologists by appearing to learn from each other.

Crickets that are made to share a cage with predatory spiders will hide under leaves to avoid attack. In experiments led by Isabelle Coolen of the National Centre for Scientific Research in Tours, France, crickets that had not been exposed to spiders were found to adopt this hiding behaviour when mixed with trained crickets. This suggests that the insects are capable of social learning — a phenomenon that, in insects, researchers have only previously observed in species that live in colonies, such as bees, ants and termites.

NEUROSCIENCE

Total recall

Nature Neurosci. doi:10.1038/nn1595 (2005)

The brain regions that help to store images and those that monitor the formation of memories have been teased apart by Yun-Ching Kao of Stanford University, California, and her colleagues.

The team used functional magnetic resonance imaging to scan the brains of 16

people as they were shown pictures of indoor and outdoor scenes. While looking at the pictures, the subjects were also asked whether they thought they would remember the image.

Activity in some regions of the brain, such as the medial temporal lobe, correlated with recall of the scene. But another area — the ventromedial prefrontal cortex (VMPFC) — showed activity when the subject predicted that they would remember the image. This, say the authors, supports the idea that the VMPFC is involved in judging the success of learning processes elsewhere in the brain.

CHEMISTRY

Born slippery

J. Phys. Chem. B doi:10.1021/jp053930j (2005)

Room temperature ionic liquids have the potential to replace more volatile organic solvents in chemical processes. But they tend to be quite viscous, making lab operations such as filtration difficult.

Now Hideaki Shirota and Edward Castner, of the State University of New Jersey in Piscataway, have found that replacing a carbon atom in the cation of an ionic liquid with silicon can make the liquid a much slipperier customer, reducing its shear viscosity by up to 7.4-fold, depending on the partner anion. Charge-density calculations suggest that this happens because polarization of the silicon-carbon bonds spreads the cation's positive charge throughout the molecule, weakening the electrostatic attraction between the ions.



I. COOLEN

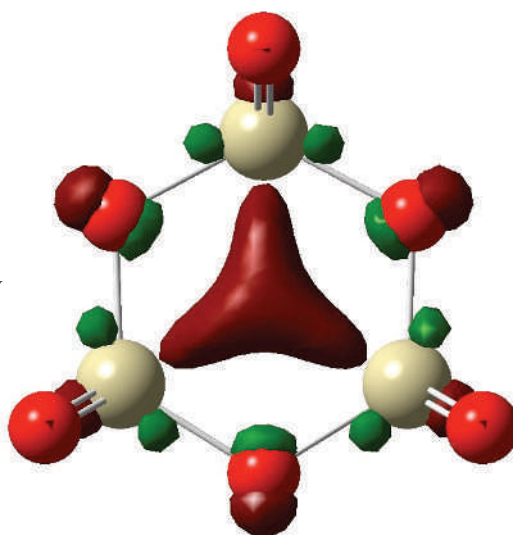
INORGANIC CHEMISTRY

Full circle

Angew. Chem. Int. Edn **44**, 7251–7254 (2005)

Aromatic organic compounds, originally noted for their smell, are also distinguished by their relative inertness. This arises because in the rings of atoms that they contain, overlap of the atoms' dumbbell-shaped *p* orbitals creates a perfectly filled — and therefore stable — electron shell in the molecular orbital.

But such aromaticity turns out not to be the sole preserve of molecules that can link up their *p* orbitals. Xin Huang and co-workers at Washington State University in Richland show that two anionic metal oxide clusters containing hexagonal rings are aromatic thanks to the overlap of the transition metal atoms' *d* orbitals (pictured right). The formulae of these metal oxide clusters are $[M_3O_9]^-$ and $[M_3O_9]^{2-}$, where M is tungsten or molybdenum



that V3 helps the virus to snare the co-receptor. The protruding nature of V3 also makes it accessible to antibodies — perhaps explaining why the immune response often targets this structure.

NEUROBIOLOGY

A charged finding

Cell **123**, 463–475 (2005)

The movements that open voltage-dependent ion channels — which are essential to signalling in neurons — may be more dramatic than some previous studies have suggested.

Measuring these movements is notoriously tricky. But a new technique has revealed that the channel protein's voltage-sensing portion, known as S4, moves 1.5–2 nanometres in response to changes in cell-membrane voltage.

The result comes from Roderick MacKinnon of the Rockefeller University in New York and his colleagues. They stuck molecules of known length to various parts of a potassium channel, then used these tethers as molecular rulers to measure changes in the channel's conformation.

CANCER

Fusion products

J. Cell Biol. **171**, 493–503 (2005)

Some types of virus might promote tumour growth by fusing cells together, hints an *in vitro* study.

Yuri Lazebnik and his colleagues at Cold Spring Harbor Laboratory, New York, examined what happens when a monkey virus fuses two cells together. As expected, in most cases the cells died. But the researchers found that the product cell proliferated if one of the fusing cells had a genetic predisposition to cancer, such as a mutation in the *p53* gene. It seems that the fusion pushes the cell into becoming cancer-like.

The findings should spark studies of the phenomenon *in vivo*. At the same time, the results sound a note of caution for experimental stem-cell therapies that are based on cell fusion.

MATERIALS

Pore show

J. Am. Chem. Soc. doi:10.1021/ja0552601 (2005)

The large surface area of porous materials can make them useful catalysts. But such activity is impaired when manufacturing techniques leave the pore walls in an amorphous, disordered state.

Hideki Sakai and colleagues from the Tokyo University of Science, Japan, report progress. They have synthesized porous titanium dioxide with crystalline walls. Titanium dioxide is widely used in solar cells and as a photocatalyst for the destruction of toxic organic materials, because of its strong oxidizing ability under ultraviolet light. The team manufactured the material from a sol-gel mixture of the precursors at 60 °C.

VIROLOGY

On the hook

Science **310**, 1025–1028 (2005)

A molecular hook that juts out from an HIV envelope protein may help this virus to gain entry to its target cells, new experiments have suggested.

HIV latches on to T cells by binding to one of their CD4 receptors and then to a co-receptor to initiate fusion with the cell. The hook, which is formed by the third variable region (V3) of the envelope glycoprotein gp120, was identified using X-ray analysis.

The team that carried out the work, led by Richard Wyatt and Peter Kwong of the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, suggests

JOURNAL CLUB

Ariel Darvasi

The Hebrew University
of Jerusalem, Israel

**A geneticist marvels at the
wealth of genetic markers that
are now available.**

More than a decade ago, I dreamed of mapping genes in a future when the availability of genetic markers was not a limiting factor.

Assuming this time would come, I worked out theoretically how I might trace the complex

traits of an organism to their genetic causes. But this approach was deemed ambitious by some, at a time when we had little sequence information.

Since then, the number of markers we have for many genomes, including that of humans, has soared from a few hundred to millions. Most of the new markers are single-nucleotide polymorphisms (SNPs), in which the DNA sequence of some fraction of the population differs by a single base.

We now have enough

information for the most demanding, and most promising, kind of genetic study: association-based whole-genome scans.

A pioneering example of such a study comes from Josephine Hoh, of Yale University in New Haven, Connecticut, and her colleagues (R. J. Klein *et al. Science* **308**, 385–389; 2005). They took advantage of chip-based technologies to study 100,000 SNPs in 96 people with age-related macular degeneration (AMD) — a disease causing poor eyesight — and in 50 healthy controls.

The scan associated the gene that encodes complement factor H with the disease. This protein regulates parts of the immune system, consistent with the idea that AMD is linked to inflammation in the retina.

As far as I know, this is the first association study to be done on such a grand scale, and many others are under way. This approach may not allow us to identify all the genes responsible for a trait — for example, some genes may have too modest an effect — but it's a promising place to start.

NEWS

Stem-cell brothers divide

Stem-cell researchers worldwide have been dismayed and confused by the abrupt end of the highest-profile collaboration in their field.

Over the weekend, the University of Pittsburgh's Gerald Schatten suddenly accused Woo Suk Hwang of possible irregularities in the donation of eggs used for his research. Schatten cut all ties to Hwang and his team at Seoul National University in South Korea. Researchers in the field, many of whom were considering collaborating with Hwang through his World Stem Cell Hub, say their plans are now on hold.

Hwang and Schatten's 20-month collaboration produced two landmark papers — the first examples of patient-specific human embryonic stem cells (W. S. Hwang *et al. Science* **308**, 1777–1783; 2005) and the first cloned dog (B. C. Lee *et al. Nature* **436**, 641; 2005). But Schatten's accusations relate to the earlier work that shot Hwang to fame, in which he established the first stem-cell lines from a cloned human embryo (W. S. Hwang *et al. Science* **303**, 1669–1674; 2004).

In a statement issued by the University of Pittsburgh, Pennsylvania, on 12 November, Schatten said: "Yesterday information came to my attention suggesting that misrepresentations might have occurred relating to [egg] donations. I have...accordingly suspended my collaboration with Professor Hwang."

Hwang has yet to respond formally. On Tuesday, he e-mailed *Nature* stating: "I have begun an investigation to find out what happened. I will inform you of the result as soon as it comes."

In May last year, an article in *Nature* presented claims that Hwang's procedures were ethically tainted by the use of eggs from two junior members of his lab (*Nature* **429**, 3; 2004). One of them, a graduate student of Hwang's, told *Nature* that she had donated eggs for the lab's research at MizMedi Hospital in Seoul. She later retracted her statement.

Egg donation is a painful and invasive procedure that requires multiple hormone injections. Donation by junior researchers is ethically suspect, because it raises the possibility that senior researchers could pressure graduate students into undergoing the procedure.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Fractured: Woo Suk Hwang (left) and Gerald Schatten's close partnership has split over ethical concerns.

Last week, Korean newspapers reported that Sung Il Roh, a fertility doctor at MizMedi Hospital who has worked with Hwang since 1995, had used illegally traded eggs to treat infertile couples. Roh refused *Nature's* request for an interview, but has so far denied that any such eggs were used for Hwang's experiments.

Hwang himself has continuously denied using either illegally traded eggs or eggs from graduate students. And Schatten had staunchly defended him. In May 2005, Schatten told *Nature* that the internal review board overseeing Hwang's research "had made it clear that no students have ever donated".

Support crumbles

Schatten's sudden statement regarding possible ethical irregularities in Hwang's work, and the fact that he has withheld related details, have shaken the world of stem-cell research. Jose Cibelli of Michigan State University in East Lansing was a co-author on the 2004 paper. He says he will await clarification of the allegations and a formal response from Hwang. But he is worried about the field. "This is a setback," he says. "We are all very confused."

Kevin Eggan, a developmental biologist at the Harvard Stem Cell Institute in Cambridge,

Massachusetts, says the allegations are difficult to weigh, because they come without any public explanation of their origins. "We have to wait and see, because there is no evidence proving them."

This has led to calls that Schatten present the reasons behind his charge. "At this stage, it is Dr Schatten's and others' responsibility to come forward with evidence," says bioethicist Insoo Hyun of Case Western Reserve University in Cleveland, Ohio.

But the allegations themselves are sufficient to cast doubt on the future of the World Stem Cell Hub. This network, launched last month, is meant to serve as a stem-cell bank and a research facility where scientists from any country can study cell lines created to order by South Korea. Many scientists in the United States, Britain and elsewhere had planned to participate, but are reconsidering their position.

Eggan, for example, says his plans are now on hold. "The allegations are serious and would have to be completely resolved for us to continue to consider collaborating with them."

And the Korean media is starting to criticize Hwang, who until now has been treated like an idol. An editorial in *JoongAng Ilbo* calls on Hwang to "prove that cloning is clean".



THE NATURE PODCAST
Listen to the stories behind
the hottest research in our
weekly show, available free
at [www.nature.com/
nature/podcast](http://www.nature.com/nature/podcast)

H.-G. LEE/GETTY IMAGES

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Japan's embryo experts beg for faster ethical reviews

TOKYO

Zealous review committees are crippling Japanese research on human embryonic stem cells, according to a plea about to be lodged at the country's science ministry.

Japan is a world leader in embryonic stem-cell research involving mice and monkeys, but work involving human cell lines is another matter. That is because review committees regularly take far longer to approve such projects than other countries do, researchers charge.

Norio Nakatsuji of Kyoto University will send anecdotal data about Japan's lag to the science ministry this month, with a request to simplify the system. "We cannot wait long, because already Japan is greatly behind other countries," he says.

In 2001, Japan decided to allow research on human embryonic stem cells and issued guidelines for researchers. Three years later, Nakatsuji, who created all three of Japan's current cell lines, began distributing them. Yet so far only 15 laboratories in Japan work on human embryonic stem cells. In a rough survey of the scientific literature for 2004 and 2005, Nakatsuji found that, of 259 papers with titles mentioning human embryonic stem cells, only three had Japanese first authors. The United States had 90 first authors and Britain had 25.

Scientific expertise with stem cells does not seem to be the problem. Japanese researchers account for more than a quarter of first authors on the 204 papers involving mouse embryonic stem cells, and more than a third of the 21 involving primates that Nakatsuji looked at. There is also no lack of funding — for example, the Ministry for Economy, Trade and Industry recently promised annual sums of ¥250 million (US\$2.1 million) for the next five years to investigate the use of human embryonic stem cells in clinical work.

Nakatsuji blames the review process, which requires approval first by an institutional review board (IRB) and then by the science ministry.

Last month he sent an informal questionnaire to 20 researchers, all of

whom complained of the time-consuming approval process, which averaged 12.5 months. Most vexing, it seems, were the questions about researchers' personal beliefs. "The boards want to know exactly how important you think the cells are. It's as if they have a soul, but they are just a bunch of cells in culture," Nakatsuji says.

"The people on the IRB seem to think of the cell lines as just as sacred as the embryos used to establish cell lines," adds Issei Komuro, a cardiovascular specialist at Chiba University. It took a year for his review board to tell him that he needed to exhaust all relevant mouse and monkey studies before moving on to humans. Eventually he gave up his plans and

decided to stick with mice.

Researchers contacted in other countries, where there is usually only one level of approval, say it generally takes a fraction of the time — two to three months in Singapore, South Korea,

Australia and Britain. In the United States, despite its reputation for restrictive policy, approval to work with permitted stem-cell lines can take as little as a few weeks.

Officials at Japan's science ministry say they are trying to improve the system. "Ministry guidelines were not clearly communicated to the institutional review boards," says Yasuhiko Ishii, director of the ministry's bioethics and biosafety office. At a meeting on 11 November he says officials discussed ways to streamline the process. They also decided to make it clear that thorough monkey studies are not always necessary before moving to human cells. "We don't want to be an obstacle to research," says Ishii. "But we need to know that proper ethical considerations are being made."

Meanwhile, Nakatsuji, who is co-hosting an international symposium on embryonic stem cells in Kyoto this week, is undeterred. He plans to create ten more human embryonic stem-cell lines in the next year, and to build a cell-processing centre to produce clinical-grade lines in the next five years. ■

Korean scientists are calling for an independent investigation, but it is not clear whether this will happen (see page 257). In the meantime, researchers are left wondering what caused Schatten's sudden change of heart.

Until recently, Hwang and Schatten had been getting on famously. "They seemed as close as they could be," says Hyun, who spent this summer studying the Korean team's ethical practices. "Gerry kept referring to Dr Hwang as his brother, and Dr Hwang's public toast to Gerry at a formal dinner was so effusive, it was almost embarrassing."

Eggan adds that just last week the two were as chummy as ever at a conference in New York. "They seemed to have every intention of continuing to collaborate in the future," he says.

Evan Snyder, a neuroscientist from the Burnham Institute in La Jolla, California, says that he received an e-mail from Schatten just before he issued his statement. "Whatever prompted this he found so exceedingly disturbing, he could not sit on it," says Snyder. "You have to realize this is a major part of his research programme as well, so to do something this precipitously, it must have been terribly shocking."

David Cyranoski and Erika Check

US budget yields scant research rises

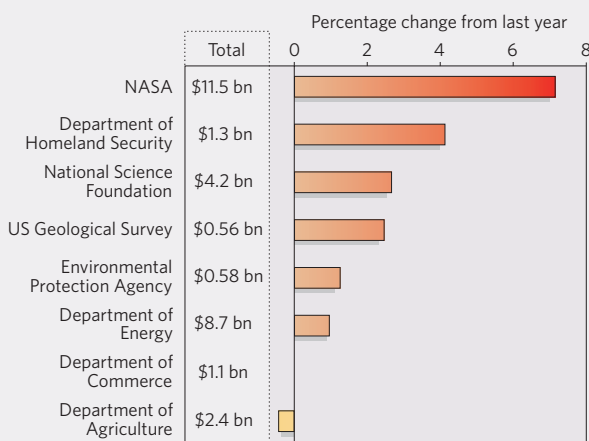
WASHINGTON DC

Key US science agencies held on to slim gains this month as Congress approved next year's budget. But they may lose these rises and more in an across-the-board 2006 spending cut to help pay for hurricane relief in the southern United States and for the ongoing war in Iraq.

Lawmakers base the budget on what the president requests in February of each year (see *Nature* 433, 559–560; 2005), but they have the authority to increase or cut funding and to specify how dollars will be spent.

The National Science Foundation has done relatively well. Last year, the agency's budget was cut by 3%, but this year it was restored by the same amount, to US\$5.6 billion, with \$4.2 billion for research (see graph). Given the tight fiscal environment, this was a reasonable achievement, says Samuel Rankin, who chairs a lobby group, the Coalition for National Science

US RESEARCH BUDGET 2006



Funding. "Under the circumstances, I'm quite pleased," he says.

NASA also did fairly well. Congress gave the space agency \$16.4 billion — nearly as much as the president asked for and 1.3% more than last year. The research budget gets a boost

of 7.3% to keep the early development of launchers and a crew vehicle for lunar expeditions on track. A shuttle mission to repair the Hubble Space Telescope is still in the picture. And lawmakers added money for several projects that the White House had short-changed, including the Space Interferometry Mission to search for planets around other stars and an Earth-science mission known as Glory.

Still, the chairman of the House Committee on Science, Sherwood Boehlert (Republican, New York), warns of trouble ahead as NASA gears up to send astronauts back to the Moon. "A renaissance costs

money, and I don't see any Medicis waiting in the wings to underwrite NASA," he said in a 3 November hearing. "There is simply not enough money in NASA's budget to carry out all the tasks it is undertaking on the current schedule."

Small conferences pay their way

SILVERTHORNE, COLORADO

The time and money spent attending small scientific meetings is more than paid back through accelerated research, suggests a survey by a conference organizer.

"The presumption is that meetings are beneficial, but the actual data to say that something positive happens are pretty scarce," says James Aiken, president of Keystone Symposia, the non-profit meetings organization in Silverthorne, Colorado, that carried out the survey.

Researchers who attended Keystone symposia on molecular and cell biology held during 2004 and 2005 later saved six weeks of research time and US\$6,000 in funding, according to median figures from the survey. The data represent a rare attempt to quantify just how effectively small meetings spur research.

The survey included only Keystone's own meetings, an admittedly transparent attempt to justify their worth. But independent scientists and organizations say that the findings could apply to other conferences as well.

"It quantifies something that we've always believed," says Howard Garrison, director of public affairs at the Federation of American Societies for Experimental Biology in Bethesda, Maryland. The federation puts on six large conferences each year that attract a total of around 35,000 participants.

The Keystone study surveyed 1,013 participants from ten conferences, ranging from obesity to the biology of hypoxia. Participants were asked how strongly they agreed with certain statements, such as: "I will leave this conference planning to

accelerate publication of some of my data." Nine months later, an e-mail follow-up confirmed whether they had acted as planned.

At the nine-month mark, 90% of scientists said they had shared information with colleagues at their

"The presumption is that meetings are beneficial, but actual data are scarce."

home institutions who had not attended the conference; 60% had established collaborations or shared information with fellow participants; and almost half had accelerated data publication. Roughly two-thirds of attendees said they had altered the direction of their research based on what they learned at the conference.

At the meetings, 85% of

scientists anticipated that information they gathered would save them time and money in the lab, but only 42% later reported experiencing such savings. And the savings estimates ranged widely — from 1 week to 2 years and from \$50 to \$2.5 million.

Meeting attendees said that conferences that are small and highly interactive, such as the Keystone symposia, have particularly high pay-offs compared with larger, more impersonal meetings.

"The lab can jump ahead by being made aware of new technologies and databases," says Charles Shoemaker, a parasite researcher at Tufts University in North Grafton, Massachusetts.

Pathologist Anjana Rao, at Harvard Medical School in Boston, Massachusetts, organized a 2004 Keystone symposium on cell

Physicists, meanwhile, are livid over Congress's treatment of the Department of Energy's Office of Science. The office provides the lion's share of universities' physics research funding and maintains several large facilities.

The office's research and development budget did creep upwards 0.6% to \$3.4 billion, but much of that money will go to congressionally mandated projects, such as a super-computing centre for Oak Ridge National Laboratory in Tennessee. As a result, several core scientific fields funded by the department face cutbacks.

Hardest hit are the Relativistic Heavy Ion Collider at Brookhaven National Laboratory in Upton, New York, and the CEBAF accelerator at the Thomas Jefferson National Accelerator Facility in Newport News, Virginia. Facing an 8.4% cut, the facilities may have to slash their operating time by up to 60%, warns Michael Lubell, director of public affairs at the American Physical Society. "Next year the department may have to seriously consider closing one of the two labs," he says.

The National Oceanic and Atmospheric Administration fared a little better, seeing its research budget rise by 2.7% to \$668 million. But most of that increase will be earmarked for an Alaskan fisheries programme. The National

Institute of Standards and Technology saw its research budget drop by the same percentage, to \$448 million. And its programme to fund high-risk research barely survived White House attempts to eliminate it, as the research budget for the Advanced Technology Program was slashed 43% to \$65 million.

In July, Congress passed budgets for agencies including the Environmental Protection Agency and the US Geological Survey — both of which got tiny research increases that do not compensate for inflation. And October saw Congress slow the Department of Homeland Security's research budget to an increase of 4.1% after several years of explosive growth.

Lawmakers will now turn their attention to finishing the final budget bills, which will set funding for agencies such as the National Institutes of Health and the Department of Defense.

On top of these numbers, all federal agencies are expecting an across-the-board cut of 2% or more to help pay for military and disaster spending. "This is not the final number," says Caroline McGuire, of the lobby group Lewis-Burke Associates in Washington DC. "There's another shoe to drop."

Geoff Brumfiel and Tony Reichardt

"This is not the final number. There's another shoe to drop."

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

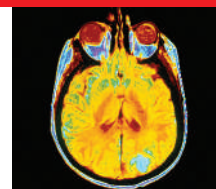
Take a view: are small conferences worth going to? One organizer says attendees save weeks in the lab.

signalling. She notes that a team of researchers might save up to \$20,000 by learning that a group at another laboratory had already developed a mouse

model that could be useful in their research.

Researchers say it is difficult to quantify the exact benefits accrued by conferences. "How

do you add all of it up to a precise number?" asks Shoemaker. "You can't. But it's big, and it's real." **Kendall Powell**



NEUROSCIENCE MEETING

Read news and reports from the Society for Neuroscience meeting in our online blog.

www.nature.com/news/columnsandblogs

Day of judgement for intelligent design

WASHINGTON DC

Two states chalked up radically different results last week in local voting over the teaching of evolution in their schools. Darwin's theory was being pitted against intelligent design — the concept that an intelligent creator shaped the course of evolution. Evolution lost one battle and won the other.

On 8 November, the Kansas State Board of Education narrowly approved a set of standards for science teaching, backed by supporters of intelligent design, that highlight “gaps” in evolution theory. “I think this is a huge victory for students in Kansas,” says Casey Luskin, a programme officer in policy and legal affairs at the Discovery Institute, an intelligent-design think-tank in Seattle, Washington.

On the same day, in Dover, Pennsylvania, district elections ousted eight of nine school-board members who support intelligent design. Last year, the board brought in a requirement that, at the start of biology lessons, teachers read pupils a statement criticizing evolution and endorsing intelligent design. A group of parents sued the school district over this, and a federal judge is now deliberating on the case (see *Nature* 437, 607; 2005).

As science educators in Kansas and Pennsylvania respond to the situation (see ‘Science advocates tackle fallout from school-board votes on evolution’), the election results underscore the highly decentralized nature of US education. Unlike most European countries, which set their curricula through a central education ministry, the United States has no established national standards, says Jay Labov, a senior adviser for education at the National

Eight members of Dover's school board lost their seats to supporters of the teaching of Darwin's theory.

Academies in Washington DC. Elected education boards set the standards in each state, and local school districts — each with their own elected board — determine how the standards are implemented in the classroom.

In the case of science, most states have voluntarily adopted a set of guidelines laid out by the National Academies in 1996. All but one of the 50 US states (Iowa) have adopted the standards wholly or in part, says Labov, and their 16,000 school districts decide how to teach them. “Under those circumstances you can imagine that such a document is treated in different ways,” he says. Such was the case in Kansas, where the academies-approved guidelines were altered to open the door for intelligent design.

A recent analysis shows how differently the teaching of evolution is treated across the United States. *Education Week*, a national journal for teachers and educators, reported in its 9 November issue that most states mention evolution in their scientific standards, but surprisingly few specify key evolutionary concepts. Only 22 states mandate the teaching of mutation and natural selection, for example. Four states fail to mention evolution at all.

“It's disappointing but not surprising,” says Wayne Carley, executive director of the National Association of Biology Teachers in Reston, Virginia. “We as science educators bear some of the responsibility for this.” ■

Geoff Brumfiel

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

AP/C. KASTER

Science advocates tackle fallout from school-board votes on evolution

In Kansas, Jack Krebs is getting ready for a fight. Krebs, a high-school maths teacher who is vice-president of Kansas Citizens for Science, fought hard against the science standards adopted last week by the state education board. He fears that their approval will encourage local districts to move away from teaching evolution.

“This vote is going to open the door for anyone who's leaning towards creationism,” Krebs says. Researchers and educators, he argues, must now focus on next

year's board elections. If more science-friendly candidates are elected, the standards could be changed before they are implemented in 2007.

Meanwhile, in Pennsylvania, Bryan Rehm is working to patch up a divided community. Rehm is a high-school physics teacher and one of the parents who sued Dover's school board last year over its requirement for teachers to read out a disclaimer on evolution in biology classes. But now Rehm has been elected to the board as

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Jack Krebs fears that Kansas schools may turn against teaching evolution.

part of a coalition that promises to find a way forward.

The issue has split the small

town of Dover. Two atheist cable-television shows have made sport of religious members of the community. Rehm, a practising Christian, has received dozens of angry e-mails from locals who support intelligent design. “What we need to do in Dover is have a conversation about science, religion and whether there is a conflict between the two,” he says.

The new board is considering introducing a comparative religion course that would discuss, among other things, intelligent design. **G.B.**

AP/O. WAGNER

Atomic agency launches bid to bank nuclear fuel

LONDON

Mohamed ElBaradei, director-general of the International Atomic Energy Agency (IAEA), has taken a step towards persuading countries to relinquish control of their nuclear fuel.

On 7 November, ElBaradei announced that the United States and Russia have agreed to contribute to an international fuel bank. The move paves the way for the first stage in a programme to bring the entire nuclear-fuel cycle under multilateral control, says ElBaradei, who, along with his agency, won this year's Nobel Peace Prize. In theory, the bank will help dissuade nuclear-hungry nations from developing facilities to enrich uranium.

Enrichment technology can also be used to develop weapons-grade material and is at the heart of current tensions between the IAEA and Iran.

Under the fuel-bank programme, nations that meet certain security standards, such as observing non-proliferation treaties, would be guaranteed a supply of fuel. Non-proliferation experts are applauding the idea of the bank, but caution that ElBaradei's long-term aims are unrealistic in today's political environment.

So far, the United States has said it will supply more than 17 tonnes of highly enriched uranium, taken from dismantled nuclear weapons. This will be 'downblended' to create reactor fuel. Russia has not said how much it will contribute.

The bank is meant to build confidence in nuclear security, says Geoff Shaw, policy adviser to ElBaradei at the IAEA's Geneva headquarters. If IAEA members can agree to the creation of the bank — perhaps when the agency's governors meet next March — they could be open to considering further parts of ElBaradei's non-proliferation plans, Shaw adds.

But such ambitions are not easy to realize. ElBaradei's second proposal is a moratorium on the development of technology for uranium enrichment and reprocessing, which can be used to recycle nuclear fuel and create weapons-grade material. Nations with this ability, such as the United States, Russia, France and Britain, would be allowed to retain it. But others would agree not to develop it in the next ten years, despite the fact that non-proliferation agreements allow them to do so for civilian purposes.

Sceptics point out that there are few incentives to sign up to the moratorium. "Are we going to say to other countries that they have to forgo their right to develop the fuel cycle?" asks Lawrence Scheinman, a nuclear-policy expert at the Monterey Institute of International Studies in Washington DC, and an adviser to three former US governments.

There are a number of reasons why nations might reject the moratorium, he says. Uranium suppliers such as Australia and Canada might find it profitable to enrich fuel before

"Supporters accept that there is a very long diplomatic fight to come."

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Mohamed ElBaradei wants to see the whole of the nuclear-fuel cycle under multilateral control.

selling it, for example. Other countries will want fuel-cycle capability so that they can at least keep the option of developing weapons. What would make Iran join the scheme, asks Mark Fitzpatrick, a non-proliferation expert at the Institute for International Strategic Studies in London. "They want enrichment for their

Scheme to track greenhouse gases takes to the air

TOKYO

Passenger planes in Japan are being pressed into service to monitor greenhouse-gas levels in the atmosphere. On 5 November, Japan Airlines flew its first plane equipped with a device that continuously measures atmospheric carbon dioxide.

Targeting flight paths from Tokyo to southeast and east Asia and to Europe, researchers say that the measurements collected by the project will provide much-needed



Plane sailing: Japan Airlines plans to provide continuous data on CO₂ levels.

information about CO₂ emissions over Asia. Eventually five planes will carry the equipment on routine flights, measuring CO₂ from the

moment they take off to when they land. As they criss-cross the region, they will build up a fuller three-dimensional picture of CO₂ than

can be obtained by ground-based or satellite observations, researchers say.

"We were looking for a way to observe carbon dioxide continuously, in broad areas and at a low cost," says Toshinobu Machida, an atmospheric researcher at the National Institute for Environmental Studies in Tsukuba, Japan. To get the project up and running the Japanese government has provided ¥80 million (US\$670,000) per year since 2003. Running costs from next year

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

L. GILLIERON/KEYSTONE/AP

nuclear weapons programme," he alleges.

But if the second stage of ElBaradei's plan is tough, the third and fourth are truly ambitious. Part three would bring the reprocessing of spent fuel under multilateral control, perhaps at a series of dedicated regional facilities. Finally, existing enrichment facilities would come under international ownership. Countries that rely on such facilities to fuel their nuclear arsenals are extremely unlikely to

agree, say experts. "For the United States, that's a dream," says Fitzpatrick.

He acknowledges that the fuel bank, if it can be made to work, would be a useful step towards achieving at least some of ElBaradei's plans. But even supporters accept that there is a very long diplomatic fight to come. "No one is naive on this point," says Shaw. "The longer term will be much more difficult."

Jim Giles

are expected to be half that.

The latest project follows on from a similar idea in the 1990s. In 1993, two planes run by Japan Airlines began carrying simple equipment to collect air samples for analysis in the laboratory. But samples were taken only twice a month on flights between Tokyo and Australia.

When the planes came up for retirement, scientists began to develop a device that could offer continuous monitoring. The equipment samples air from the front of the plane's engines and so does not pick up

the aircraft's own emissions, says Yukio Nakagawa, manager at the engineering department of Japan Airlines. The hardest task, he notes, was creating a device with the appropriate specifications given the limited time and cost.

Inside the plane's cargo compartment, air flows through spectrometers that continuously measure the CO₂ concentration. Associated equipment detects other greenhouse gases, such as sulphur hexafluoride.

Although the equipment has

so far flown on just one plane, the company plans to add devices to four more of its Boeings by the end of next year.

Toshihiro Ogawa, a retired atmospheric chemist formerly at the University of Tokyo, says the project should help researchers to quantify carbon dioxide emissions and so make it easier for countries to conform to the Kyoto Protocol on climate change. "Figuring out the real carbon dioxide emissions per country is our big homework," he says. ■
Ichiko Fuyuno

ON THE RECORD

"I'd like to say to the good citizens of Dover: if there is a disaster in your area, don't turn to God. You just rejected him from your city."

Christian talk-show host Pat Robertson attacks the result of the school-board election in Dover, Pennsylvania, which saw eight proponents of intelligent design lose their seats (see page 267).

"Even if a dynamic physics model suggests cow tipping is possible, the biology ultimately gets in the way."

Margo Lillie of the University of British Columbia argues that, contrary to popular belief, it is far from easy to tip over a cow.

Sources: Reuters, The Times

SCORECARD



Music

Astronauts on the International Space Station get an unusual wake-up call as Paul McCartney becomes the first person to broadcast live music into space.



Espionage

A Russian researcher claims to have found a way to control turtles remotely. The creatures can be used to spy behind enemy lines and even deliver bombs, he says.



Lunar land prospecting

Chinese authorities have shut down a 'Lunar Embassy' that sold plots on the Moon for US\$37 an acre.

OVERHYPED

Treatments for bird flu

The threat of an avian flu pandemic has worried many people. As a result, a number of unconventional remedies are being offered as ways to combat the disease. In recent weeks, oil of oregano, colloidal silver and the pickled Korean cabbage known as *kimchi* have all been touted for their purported flu-fighting properties. None has been proved to work — and none can beat the advice of the experts in the event of a pandemic: just stay at home.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

The Dalai Lama talks to neuroscientists about the shared goals of science and religion.

Dalai Lama speaks at neuroscience meeting

Despite earlier protests that he was not qualified to speak at a scientific meeting, the Dalai Lama met a warm reception at the Society for Neuroscience meeting in Washington DC on 12 November.

In a keynote lecture, the Tibetan religious leader told 14,000 scientists that science and religion have similar goals — seeking truth with an open mind. He said that Buddhist teaching should change if it conflicts

with the results of modern science.

His speech had been opposed by about 800 people who signed a petition asking that he not speak at the meeting. Six scientific abstracts were withdrawn in protest, according to the society. But only one protester showed up outside the press conference room: neuroscientist Pei Wang, of the State University of New York at Buffalo, had scrawled 'Dalai Lama not qualified to speak here' on her conference bag.

Cosmic-ray observatory obtains first results

Astronomers are celebrating the arrival of the first results from the Pierre Auger Observatory, a vast network of cosmic-ray detectors that will cover about 3,000 square kilometres in rural Argentina.

More than half of the observatory's 1,600 detectors are now working, and an international group of researchers met on 10–11 November for a dedication ceremony in Malargüe, Argentina. The new data include a spectrum of cosmic rays at the highest energy levels, and the first results from a search for highly active cosmic-ray sources. "Understanding the origin of these particles should open up a lot of different

questions," says Angela Olinto, an astronomer at the University of Chicago in Illinois who works with the observatory collaboration. "But we have to start by knowing where they come from." The observatory is due to be completed next year.

Nobel physicist goes to prison after car crash

A Nobel laureate was sent to prison last week for two years after his high-speed driving caused a fatal crash in California.

John Robert Schrieffer, who shared the 1972 Nobel physics prize for helping to describe superconductivity, was sentenced in Santa Maria on 7 November after being convicted of vehicular manslaughter. Schrieffer, 74, is on unpaid leave from the US National High Magnetic Field Laboratory in Tallahassee, where he is chief scientist.

In September 2004, Schrieffer drove a sports car at more than 100 miles per hour into a van. His driving licence was invalid at the time, having been suspended because of prior speeding violations.

"Schrieffer is extremely remorseful," says his attorney, Roger Lytel. "He liked to drive fast and it caught up with him." He faces a civil trial in January for damages.

World's largest iceberg splinters into shards

In the last week of October the Envisat satellite of the European Space Agency captured radar images of the colossal Antarctic iceberg B-15A splitting into knife-shaped shards. Agency officials think that the iceberg had run aground, causing stresses that snapped the ice along existing faults such as crevasses.

"This kind of break-up is not the 'death knell' for the iceberg, but rather part of its 'adolescence'," says Douglas MacAyeal of the University of Chicago in Illinois.

MacAyeal is studying the iceberg from his base at the McMurdo research station on Ross Island, near to where it is drifting.

B-15A, with an area of about 2,500 square kilometres, was the largest surviving fragment



of an even bigger iceberg that calved from the Ross Ice Shelf in March 2000. While in the McMurdo Sound, it has at times been an obstacle to boats and has restricted penguins' access to their feeding grounds at sea.

Research centre refuses tobacco-company funding

Germany's largest health research centre is the first research institute in Germany to adopt an ethical code that bans all kinds of funding from tobacco companies.

The code bans the 500 researchers at the German Cancer Research Centre (DKFZ) in Heidelberg from receiving any financial

support, including research funding, lecture fees and prizes such as the prestigious €100,000 (US\$117,000) Philip Morris research prize. "It is a binding ethical code for all our staff members," says Martina Pötschke-Langer, head of the cancer prevention unit at DKFZ.

German health researchers accepted millions of dollars of tobacco-industry funding in the 1980s and 1990s (*Nature* 435, 866; 2005 and *Nature* 410, 725; 2001).

The DKFZ itself accepted grants from the tobacco industry until its director halted the practice in 1982. Pötschke-Langer says she hopes the new ethical code "will set an example for other German health institutes".

ESA

Venus Express may deliver clues about global warming

Next stop, Venus. The Venus Express probe of the European Space Agency has begun its mission to investigate the planet after a successful launch on 9 November.

When the probe arrives at Venus in April 2006 it will study the planet's thick atmosphere, which being rich in carbon dioxide, keeps the surface at a fiery 450 °C.

Scientists say that the mission could help them to understand our own planet's global warming. "Originally, Venus and Earth must have been very similar planets," says Jean Jacques Dordain, the space agency's director-general. "We really need to understand why and how they eventually diverged."

The €220-million (US\$265-million) mission is scheduled to last for just two venusian days — but because Venus rotates so slowly, that should provide more than 16 Earth-months of observations.

Root of all evil? Park rangers drench invasive weeds with blue herbicide.



SHOOT TO KILL

The US government has adopted a tough approach to battling harmful exotic plants: specialist strike teams. But can they prevail? **Emma Marris** finds out it's not all black and white.

Chesapeake bay in Virginia played host to some of the first British colonists. And along the banks of the York River, crumpled beer cans and someone's soggy hunting cap bear witness to their success. Another colonial type has also put down roots by the river: a few feet above the high-tide line sit some feathery reeds. They are *Phragmites australis* — and they are the enemy.

This strain, thought by some to have been introduced from Europe a century or so ago, is now classified as an invasive species — one which aggressively steals habitat from native flora. Today, the reeds will come under attack from an invasive-species strike team.

Armed with high-tech gear, these specialists travel from park to park — all-terrain vehicle in tow for the off-road areas — ready to pull, poison or burn anything that is out of place. Since their inception in 2000, the National Park Service's strike teams have treated some 270,000 hectares. But success depends on the yardstick one uses: plant seeds can survive for

decades and most invasives return sooner or later, often within a few years of treatment. Some ecologists even question the rationale behind the entire approach.

Dressed in a pale khaki uniform and thigh-high waders, Kate Jensen is carefully recording the exact location of the *Phragmites* using a hand-held global positioning device, which chirps merrily when it reaches enough satellites. Half an hour later, her teammates — Dale Meyerhoeffer and Matthew Overstreet — arrive with a 200-gallon tank of a herbicide called Habitat. The team mixed the chemical early in the morning and added a fluorescent blue dye that will make it obvious which plants have been hit.

The aim, Jensen says, is to make the area liveable for native plants. She points to a handful of cattails behind the *Phragmites*. "That's what we want here. We like those."

The Park Service's 16 exotic-plant management teams were conceived as a way to spread scant resources over hundreds of US national parks. Each team covers a wide area — this one

handles the mid-Atlantic region, which stretches 550 kilometres from Pennsylvania to Virginia. The teams are made up of park employees, contractors, conservation students and volunteers, whose shared goal is to wrestle the parks back into something like their pre-colonial condition.

To this end, the mid-Atlantic team goes after tree-of-heaven (*Ailanthus altissima*) from China and autumn olive (*Elaeagnus umbellata*), an east Asian import once planted to fatten game birds. They also tackle privets (*Ligustrum* spp.) — hedgerow plants that form dense thickets and compete with natives — and mile-a-minute weed (*Polygonum perfoliatum*), whose name betrays its bad habit.

Cutting costs

The introduction of such alien plants is a growing concern worldwide, as international travel and shipping accelerate the spread of seeds and cuttings around the globe. Among the most damaging additions to the United States are leafy spurge (*Euphorbia esula*), which has rendered great swathes of western prairie useless for grazing, and salt cedar (*Tamarix* spp.), which depletes groundwater and moves salt to the soil surface. Crews cutting the cedars back with chainsaws report a

E. MARRIS

salty taste on their tongues as they work. The cost of the damage is estimated to run into tens of billions of dollars each year.

And these figures do not take into account intangibles such as loss of fresh water or recreation opportunities, says Mike Ielmini, invasive-species coordinator for the US Forest Service, based in Washington DC. "Until we get some definitive studies, it is going to be very hard to say how much it costs the United States," he says.

In the face of such economic damage, the US government's response has been increasingly aggressive. In 1999, President Bill Clinton signed an executive order that crystallized the approach to invasive species: all agencies were to avoid spreading them and to stamp them out, wherever possible. Today, the federal government spends more than \$1 billion annually controlling invasive species¹.

But is such large expenditure necessary? Although many thousands of plant species have accompanied human migrations to new lands, most manage to coexist in one way or another with what is already there. Only a small percentage are truly invasive, pushing out the native flora and altering the local ecology. And although the prevailing ethos says that all invasives should be eliminated, not all ecologists agree.

Required weeding?

Removing invasives is more a human preference than a scientifically grounded prerogative, argues Mark Davis, an ecologist at Macalester College in St Paul, Minnesota. "There is a huge amount of arbitrariness here," he says. "Can't we just forget about where they came from, identify species that are causing us problems according to our values, and then deal with them?"

Although invasives have been called the second most serious threat to endangered species after habitat loss², a recent review of the literature suggests that there are few solid data to support this idea³. Many ecologists now agree that invasive plants, in particular, seldom drive other plants extinct by competing with them directly for resources. The view that ecosystems are stable places in which every niche is filled until some disruption occurs has given way to something else. "I think that the natural world out there is more like a swirling and boiling cauldron," Davis says, where change and challenge are the order of the day.

Perhaps the fears about invasive plants are overblown? Not according to Peter Vitousek, an ecologist at Stanford University in California, who says interloping plants more often cause local extinctions by changing something fundamental about the area. "They don't just compete with or consume native species, they change the rules of the game," he says, by changing the composition of the soil, the availability of water or the frequency of fire. "It is often a subtle and long-term phenomenon."

The truly problematic invasives tend to share several characteristics that turn them

"They don't just compete with or consume native species, they change the rules of the game."

— Peter Vitousek

from, say, a charming ornamental shrub into a rampaging monopolizer of wild land. They reproduce quickly and explosively, either by prodigious production of seeds or by fast-growing rhizomes — horizontally growing roots that sprout new plants. Often, invasives are plants that in their native regions colonize areas after a fire. They do not attract local insects or diseases. And some, including tree-of-heaven and leafy spurge, exude noxious substances that knock out competing plants.

Phragmites spreads both by seeds and rhizomes to produce thick stands with up to 200 stems per square metre. Its aggressive tactics are the reason for Overstreet's herbicidal attack. He leans back, balancing the weight of the hose. On the weekends, he rides broncos competitively. He is also a veteran of park management, apt to say things like "I did that running from a bear" or "That was when I was hunting hogs in the Smokies".

Meyerhoeffer pays out more hose so Overstreet can work his way behind a ghostly stand of dead *Phragmites* from last year, which still shelters some persistent shoots. "I'm getting a few back here," he calls. "I'm leaving the rest

Phragmites australis
stands tall among
the locals.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

for job security." After a pause, he adds: "That was a joke."

The strike-team approach is winning converts among managers of public lands. The US Fish and Wildlife Service set up five of its own strike teams in 2003 to attack invasive species within National Wildlife Refuges. These are parcels of land set aside to conserve flora and fauna. Although the word 'native' does not appear in the teams' official remit, it is presupposed, says Michael Lusk, invasive-species coordinator for the refuges. The Forest Service, which manages land for public and commercial use, spends more than any of the other services on managing invasive species; it says it is planning to set up strike teams of its own.

Bush administration

For its part, the Park Service is charged with preserving places of cultural importance to the United States, such as Alcatraz and the Grand Canyon. So are the reeds around York River being killed for their historical inaccuracy? Rita Beard, who runs the plant management teams, says no. "It is not just a matter of aesthetics. It is not just a matter of saving the plant communities that have been here historically," she says. "It is also about maintaining ecosystems that can withstand the ecological changes that will inevitably occur."

The idea is that an ecosystem with lots of co-evolved species will be more resilient to changes, such as the spread of a disease or a change in climate, than a smaller group of introduced species. The assumption is that aggressive invasive species will reduce the biodiversity in an area and that this will make the area less stable. Ecologists such as Davis are not ready to sign on to this intuitive premise, however.

Davis, who sees invasive plants as less of a menace than many, worries that resources are being spent on fighting exotics that could be used for other conservation processes. "More and more people are now questioning — with data — what was originally presented as a kind of simple idea: that invasive species are inevitably a huge threat," he says.

Whether they are or not, culling these plants is not easy. For the mid-Atlantic team, the long, hard hours battling plants can feel a bit Sisyphean. "This is a never-ending process. You can get a handle on it, but the invasives are always creeping in," admits Jensen. Still, she says, the job is important, and it is clear that it colours her world view. She reaches through some reeds and snaps a bent twig off a tree. "This is native cherry. We like that."

Emma Marris is a reporter for *Nature* in Washington DC.

1. National Invasive Species Council
www.invasivespeciesinfo.gov/council/FY06budget.pdf
2. Wilcove, D. S., Rothstein, D., Dubow, J., Phillips, A. & Losos, E. *Bioscience* **48**, 607–615 (1998).
3. Gurevitch, J. & Padilla, D. K. *Trends Ecol. Evol.* **19**, 470–474 (2004).

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

The long-range forecast

The Himalayas, roof of the world, are springing a leak. As the climate warms up, melting glaciers are threatening the livelihoods of millions. **David Cyranoski** reports.

The mighty Himalayas straddle many countries and are home to a host of ethnic groups. But all along the mountain range, local communities are finding that they have at least one thing in common: their way of life is being threatened by changes to their environment.

In Nepal, rising temperatures are swelling glacial lakes to bursting point. Across the mountains in Tibet, herdsman are struggling to feed their livestock on an increasingly deteriorating landscape. In recent years, the locals have begun to blame global warming for many of their troubles — and the data now being collected suggest that they may be right.

“Local people do not know much about climate change, but they can very clearly see that there is a change affecting them,” says Lifeng Li, at the China branch of conservation group the WWF.

Billions of people rely on water that ultimately comes from the Himalayas. The range is home to the 14 highest peaks in the world, and the area's snowpack feeds the flow of several major rivers.

Hydrological models¹ suggest that the effects of climate change in the region could be far-reaching. And reports published earlier this year paint a grim picture of vanishing glaciers

and declining water supplies^{2,3}. Future disasters could include floods, droughts, land erosion, biodiversity loss, and changes in rainfall and in the annual monsoon. As a result, scientists and conservation groups are monitoring the region and asking what can be done to help its inhabitants to adapt to the changes.

Swept away

Both sherpa mountain guides in Nepal and cattle herdsman in Tibet stand to benefit from these initiatives. Climate change can have an immediate and disastrous effect on their lives. In 1985, Nawa Jigtar, a senior monk in the village of Ghat in Nepal, heard a loud noise and rushed outside, only to watch helplessly as his cows were carried away by a deluge. “If it had come at night, none of us would have survived,” he told the WWF in a recent documentary, *Meltdown in Nepal*.

That flood occurred when the Dig Tsho glacial lake burst its banks, wiping out more than a dozen bridges as well as a new hydro-electric plant. Studies by the Organisation for Economic Co-operation and Development (OECD) and others have shown that at least 20 such glacial lakes are at risk of bursting in Nepal⁴, with more in Bhutan. The lakes form naturally from the meltwater of glaciers, but as

climate change forces the glaciers into an ever-faster retreat, the excess water floods into the lakes. Fragile moraine dams — piles of rock and debris left by the retreating glacier — are all that stand between the water and the communities downstream, says Shardul Agrawala, a climate-change administrator at the OECD. Earthquakes, landslides or slope instability can trigger the natural dams to collapse.

And the problem seems set to get worse. In March, a WWF team reported on how quickly the Himalayan glaciers are melting², and what consequences that might have for water supply in India, China and Nepal. Temperatures in Nepal have been rising steadily over the past few decades, and climate-change models predict that they will rise a further 1.2 °C by 2050, with a total increase of 3 °C by 2100. As a result, the glacial lakes are growing in both number and size.

“Most glaciers are in retreat, so glacial lakes have to hold more water,” says Ninglian Wang, a glaciologist at the Cold and Arid Regions Environmental and Engineering Research Institute (CAREERI) in Lanzhou, China. Nepal's glaciers are shrinking at a rate of 30 to 60 metres per decade⁴.

On the Tibetan side of the range, 50% of the glaciers were retreating during the period 1950

to 1980 — that rose to 90% in the 1980s and to 95% in the 1990s², according to data gathered by Tandong Yao, head of the Chinese Academy of Sciences' Institute of Tibetan Plateau Research based in Beijing. The runoff from melted glaciers in China is now roughly equal to the annual flow of the Yellow River.

"Melting glaciers are one of the most convincing pieces of evidence for the impact of climate change," says Nigel Arnell, a specialist in climate resources at the University of Southampton, UK. But the picture isn't so clear when it comes to other data, such as the cause of land deterioration in Tibet.

This summer, a team from CAREERI sponsored by Greenpeace documented shrinking glaciers and lakes in the source region for the Yellow River³, which rises on the Tibetan plateau and is a crucial water resource for millions of Chinese. The group found that the region is experiencing vegetation loss of 3–10% each year. If the trend of rising temperatures continues, the team predicts a decrease in water availability of 20–40% over the next 50–100 years, and a fall in total agricultural output including wheat, rice and corn crops of 10% by 2030–50.

Down to earth

But the question of land use — particularly overgrazing — is complicating the picture. The locals tend to reject the idea that they are partly to blame for the deterioration. For instance, Tibetan herdsmen near the source of the Yangtze River, who were interviewed by WWF researchers, were adamant that global warming had ruined the amount and quality of their grasslands, as well as increased the spread of disease from animals such as rats. And one of the main findings of the subsequent report² put together by members of the Chinese Academy of Sciences, was to correct the notion that overgrazing had destroyed the ecosystem.

Most scientists, however, believe that overgrazing is playing a major role in the vegetation loss — the question is how much. The issue highlights one of the major problems in assessing the changes caused by global warming. Anecdotal surveys form the bulk of the data for these regions, and experts warn against relying too heavily on this kind of evidence. The regional studies are often not rigorous, notes Martin Parry, a senior research fellow at Britain's Met Office. "Even if you can estimate the climate change, you can look at what the impact would be on agriculture today, but agriculture in the future will be different," he says. "You need to take into account social, economic and political change."

Occasionally, local governments have taken the initiative to try to mitigate the effects of climate change. In February, the Chinese government allotted 7.5 billion renminbi (US\$930 million) for conservation projects in the source regions of the Yangtze and Yellow rivers. To solve the problem of land degradation, officials are talking about relocating many of the



Researchers interview Tibetan herdsmen (above) and measure glaciers in a bid to assess the effects of climate change on the Himalayas.

Tibetan herdsmen, says Li. He argues that such an approach would be too costly, both financially and in terms of its cultural impact. "We should try to improve productivity of the land rather than move people," Li says — perhaps by selecting more robust species of grass and improving irrigation in the region.

Peak practice

Dealing with the threat of glacial lakes is a little more straightforward — if there is money available. Beginning in 1998, the Netherlands spent US\$3 million to drain Tsho Rolpa, one of Nepal's most dangerous glacier lakes. The project lowered the water level by 3 metres, but the respite is only temporary, warns Agrawala. Meltwater is likely to fill it up again shortly, he notes.

A new book from the OECD that examines the situation in Nepal⁴ suggests other measures, such as early warning systems for floods to reduce the risk to local communities. It also proposes using smaller hydroelectric plants

spread out on several rivers, rather than relying on one major plant that could be wiped out in a single lake burst, as happened in 1985.

Such solutions are just one approach to a multifaceted problem. Regional policy-makers must juggle social and economic needs while worrying about how to adapt in the face of global warming. "There is a lot of talk of climate change, but how do you tell national planners to take it into account?" asks Agrawala.

One possible solution could be a new adaptation fund, to be considered when member states of the United Nations Framework Convention on Climate Change meet in Montreal starting on 28 November. The framework would provide funding, insurance and technology transfer to developing countries whose geography puts them at risk from climate change — such as countries prone to drought, and those with mountainous ecosystems.

Regardless of what aid it could get in the future, China may already be getting a taste of things to come. In Qumalai County, near the headwaters of the Yangtze, wells have recently gone dry and smaller rivers have vanished completely. If climatologists are correct, these water shortages may be a harbinger of the future along the Himalayas. ■

David Cyranoski is Nature's Asia-Pacific correspondent.

1. Barnett, T. P., Adam, J. C. & Lettenmaier, D. P. *Nature* **438**, 303–309 (2005).
2. WWF Nepal Program *An Overview of Glaciers, Glacier Retreat, and Subsequent Impacts in Nepal, India and China* (WWF, 2005); available at http://www.panda.org/downloads/climate_change/himalayaglacierrreport2005.pdf.
3. Ding, Y., Liu, S., Xie, C., Zhang, Y. & Wang, J. *Yellow River at Risk: An Assessment of the Impacts of Climate Change on the Yellow River Source Region* (Greenpeace, 2005); available at http://activism.greenpeace.org/yellowriver/yrs-english_web.pdf.
4. Agrawala, S. (ed.) *Bridge Over Troubled Waters: Linking Climate Change and Development* (OECD, Paris, 2005).

See also pages 283 and 285 in this issue.

Taking a stand

A protest by Chinese graduate students at Yale University has revealed the plight of a vulnerable workforce in US labs. **Geoff Brumfiel** investigates.

Last month, Xuemei Han was at her wit's end. A second-year graduate student in the Department of Ecology and Evolutionary Biology at Yale University, she was facing expulsion. Efforts to transfer to the university's forestry school had failed, and it looked as though the 26-year-old might have to return to China within a matter of weeks. "I had a lot of pressure on me, and I did not feel confident anymore," she says.

In June, Han had been told that she was "not in good academic standing" with her department — an accusation she disputed. She had passed her qualifying exams at the first attempt and, after a few more tries, her required language exam as well. So she did something that many Chinese graduate students would never dream of doing: on 20 October she filed a complaint against Yale, accusing the university of treating Chinese students unfairly. The only Chinese student in her department, Han wrote in her complaint that she suspected professors were reluctant to work with her because they thought she would need extra help preparing manuscripts and grant proposals.

Her grievance quickly gained a high profile on campus and beyond. Three other graduate students filed supporting testimonials that detailed problems they had experienced in their departments, and just over half of the 274 Chinese graduate students at Yale signed a statement backing her. The case garnered media coverage in the United States and even made the evening news in China. Within a week, university administrators relented and allowed Han to transfer to the department of forestry, where she had found an adviser willing to support her.

Yale flatly denies any accusations of discrimination against Chinese students. Department members and administration officials declined to comment on the details of Han's case for legal reasons, but in a statement, Yale spokesman Tom Conroy said: "Yale has a long standing tradition of being a welcoming and supportive university for international students, and especially those from China."

Whether or not it was discrimination, Han's



Yale students rally to highlight the problems faced by Chinese postdocs working in the United States.

GESO

story taps into a rarely seen vein of discontent among Chinese students and postdocs across the country. Chinese nationals are by far the largest group of foreign academics working in US universities. Between 1985 and 2000, some 26,500 Chinese students earned science and engineering PhDs in the United States — more than double the number of students from all of Western Europe, according to the National Science Foundation. And a recent survey of postdocs by scientific research society Sigma Xi in Research Triangle Park, North Carolina, showed that Chinese postdocs tend to work longer hours for less pay than their American counterparts (see graphic).

Culture shock

Many Chinese come to the United States to participate in cutting-edge research, but must first overcome language barriers, cultural differences and visa hassles. They frequently feel isolated from their US lab-mates. And although all graduate students are at the mercy of their advisers, foreign students are especially vulnerable. They lack alternative options, so a disagreement or funding problem is all that it takes for them to be sent back to China. "A lot

of people live in fear," says Cong Huang, president of the Association of Chinese Students and Scholars at Yale and a third-year graduate student in the statistics department.

The high percentage of Chinese in the lab is no coincidence. US researchers are happy to recruit academically gifted Chinese scholars, while the best and brightest Chinese are drawn to the country by research opportunities that they cannot get at home.

That opportunity is what brought Han from Inner Mongolia to Yale in 2003. She received her undergraduate and master's degree in ecology from Beijing Normal University, but had never travelled outside China. "Ecology research has only just started in China, so my professors recommended that I study here," she recalls. She was ecstatic when she learned that Yale had admitted her to a PhD programme with funding from a Fan Family Fellowship, which supports Chinese students.

But shortly after arriving in the United States, Han ran into difficulty. Like many Chinese students, she had studied English extensively in China, but that training focused primarily on reading and writing, not speaking. "The first semester was very hard," she

says. "In physics and other departments, there are other Chinese graduate students who can help, but I was the only one in my department."

Han's experience is not unusual. Many Chinese students have trouble fitting in when they first reach the United States, according to Hongwen Zhu, a graduate student at the Albert Einstein College of Medicine in New York. Zhu says many students are embarrassed to admit that they don't understand what is being asked of them, or they are reluctant to raise their concerns vocally with their professors. "Most Chinese students tend to be very quiet, and this is a very big problem," he says.

Han made steady progress in her language skills, but it came at a cost. She was unable to teach, a requirement of her department, and she had trouble finding a research adviser.

On the edge

Still, Han was shocked to learn in June that she was no longer in good standing with her programme. The department told her that without an adviser she would be expelled from the programme on 31 August. With help from members of the ecology department, Han eventually found someone who was willing to advise her in Yale's forestry school. But when she tried to transfer, she was informed that she would lose the Fan Family Fellowship.

Foreign students and postdocs frequently run into these sorts of funding problems, says Ji-Cheng Wang, a postdoc cancer researcher at City of Hope Hospital in Duarte, California. Unlike American students, who can switch advisers if necessary, many foreigners are financially tied to their principal investigator (PI). "If anything happens to the PI then the student is put at risk," Wang says.

This relationship can put students in a precarious position. When Wei Fu, not his real name, moved from Peking University to become a postdoc at a midwestern university, he was hoping for a chance to expand his own research career in biophysics. Instead, Fu's lab director asked him to devote most of his time to existing experiments. "I didn't have much independence, I didn't feel free," he says. When Fu told the PI of his unhappiness, he found himself suddenly out of a job. He had just three months to scramble for a new position, or risk expulsion from the country. Eventually, he managed to find a position at a lab in California. "You can imagine that I was very stressed," he says.

That stress has been exacerbated by recent US and Chinese immigration policy. Most international students and scholars get a multiple-entry visa for the duration of their studies, but Chinese students must reapply for a new visa every six months. That is an improvement over the old rules, which required students to reapply each time they left the country, but it still causes trouble for researchers such as Yangheng Zheng, a postdoc studying high-energy physics at the University of California, Los Angeles. While conducting



Speaking out: Xuemei Han addresses students protesting at her possible expulsion from Yale University.

graduate research at the University of Hawaii, Zheng frequently travelled between the United States and Japan, and each trip required a new visa both ways. "In three years I used up all of my passport's pages," he says. Although the situation is better now, there are still problems, he says. Two months ago, on his latest excursion to CERN, the European particle-physics lab, he ended up stuck in Geneva for three weeks waiting for a US security check.

Sink or swim

There is little consensus in the Chinese community over how serious these issues are. Some students and postdocs that *Nature* spoke to said they had not encountered significant problems, and many reported strong relationships with their advisers, who helped them resolve issues. "The people I know are very nice to me," says Ye Jin, a postdoc in molecular biology at the University of California, Berkeley. "When I try to write papers and proposals my PI has been very patient and corrects my grammar. She has been very encouraging."

"Language is not a barrier if you are willing

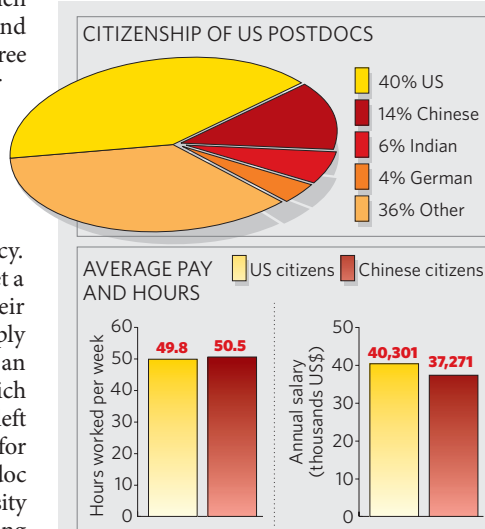
to learn," adds Grace Wong, the president of Student Vision, a Boston-based group that helps students find jobs in biotechnology. "If your skills are good and you're willing to work really hard, any boss will love you."

But Huang disagrees. "We really appreciate that the university gives us the chance to come here and study," he says. "But even if you work hard, sometimes you still have the risk of being kicked out because of a funding problem or a disagreement with your adviser." Huang's group, together with Yale's Graduate Employees and Student Organization, is now asking for modifications to the university's grievance process for international students.

Regardless of whether the issues are real or perceived, the United States has seen a decline in Chinese graduate students in recent years. First-time enrolments were down 8% in the 2003–04 school year — although they rose by 3% this year, according to the Council of Graduate Schools, a non-profit consortium of graduate educators. Heath Brown, director of research at the council, says that the decline was primarily the result of stricter visa policies after 11 September 2001, which made the United States seem less welcoming (see *Nature* 427, 190–195; 2004). Brown sees this year's reversal as "very positive", but he adds that much needs to be done to improve the image of the United States among Chinese students and scholars.

As for Han, now that Yale has allowed her to transfer to the forestry school and retain her fellowship, she says that she is "very happy". She hopes that her story will encourage others to speak up when they encounter trouble working or studying in the United States. "Most students don't know even one example in which a student fought and won," she says. "I hope that after me, more students will speak up for themselves."

Geoff Brumfiel is *Nature's* Washington DC physical sciences correspondent. See Editorial, page 258.



SOURCE: G. DAVIS/SIGMA XI

BUSINESS

Wisdom of the crowd

Decision makers, wrestling with thorny choices, are tapping into the collective foresight of ordinary people. Jim Giles reports.

Like all of their rivals in the pharmaceutical industry, executives at Eli Lilly routinely need to make tough predictions. How much of a new product do they expect to sell? Is a competing drug going to win approval first?

Millions of dollars of revenue hang on getting these predictions right. So it might come as a surprise to learn where some executives at the Indianapolis-based firm have looked for answers: they've been asking readers of *USA Today*.

Lilly is one of several major corporations now dabbling in 'prediction markets' — decision-making tools that harvest the collective wisdom of groups of ordinary people. Participants are asked to buy and sell shares in real outcomes, and are rewarded for betting on outcomes that turn out to be correct.

Advocates of the prediction markets claim that the predictions of such participants can end up being better than those made by specialists.

"The idea of prediction markets is a powerful one," explains Thomas Malone, a management specialist at the Massachusetts Institute of Technology. "They let many people contribute to the collective assessment of a future event. It's a surprisingly effective way of integrating information."

Lilly runs its markets in conjunction with NewsFutures, a firm based in New York that sells prediction-market software. When the two companies first collaborated in 2003, *USA Today* was experimenting with the idea as a game for its readers. Lilly subsequently asked 250 of these readers to make predictions about some of its business issues.

The group was invited to buy or sell shares pegged to specific predictions, such as the number of drugs that would be approved in a year by the US Food and Drug Administration. Shares in the correct prediction paid out virtual money at the end of the year, and Lilly motivated traders by stumping up

\$10,000 in real money to reward the winners.

Although there hasn't been much independent analysis of prediction markets, the data that are available suggest the concept carries promise. Markets aimed at predicting the outcome of US presidential elections and the names of Oscar winners have, according to some assessments, outperformed other indicators such as opinion polls. Emile Servan-Schreiber, NewsFutures' chief executive, says

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Safety in numbers: ordinary people can pip specialists in forecasting what's to come.

the same pattern emerged with the 2003 Lilly market; outside traders did better in their predictions than company experts.

Supporters say prediction markets motivate participants to carefully study information related to whatever they are trading in. And as outsiders, they lack the vested interest that can colour the predictions of internal staff, or even some consultants, who might, for example, have a historical empathy for a particular product line. "The markets provide an interesting counterpart to predictions from groups, such as lobbyists, who tend to see things in black and white," says Servan-Schreiber.

Since its initial study, Lilly has commissioned two further prediction markets, the second of which is running this year and using its own sales staff as participants. In this market, predicted outcomes, such as the revenue that

a specific Lilly drug will achieve in different quarters of the year, are assigned a price: a trader might buy a share in that outcome if, for example, they think the share price underestimates the outcome at the end of the quarter. Lilly declined to discuss the details of this market, but Servan-Schreiber claims that it has already outperformed standard internal forecasts.

Rival firm Abbott Laboratories, based in Illinois, has also purchased prediction-market software from NewsFutures, but declined to say how this is being used. And similar schemes are finding myriad uses elsewhere. The University of Iowa in Iowa City, for example, has set up a prediction market to help health authorities assess how the state will be affected by influenza in the coming year.

But how much trust can be placed in these tools? Much of the enthusiasm for the concept

has been driven by the publicity surrounding their use as predictors of the outcomes of US presidential elections. But the data from these are slim, and corporate leaders may be sceptical about what prediction markets are really worth.

Charles Manski, an economist at Northwestern University in Illinois, has looked into the markets' performance, and takes issue with one of their assumptions, namely that traders' beliefs alone determine market price. He thinks that the price is also affected by differences in traders' budgets and attitudes to risk, and that unless these other two influences are understood, predictions will

be misinterpreted. Manski warns against believing all the claims made about the approach. "The problem with prediction markets is that they've been hyped by the kinds of people who believe that markets solve all problems," he says.

Malone concedes that the markets still need to earn their spurs. But he stresses that information from them should at least be considered — especially when it runs counter to mainstream thinking. "At a minimum, you should do more investigation to find out why the market said what it did," he suggests. Once firms start to see the prediction market as just another tool, it'll become an everyday part of business management, he predicts. "Information technology has allowed the cost of doing this to fall to almost zero," Malone says. "It will become routine."

A. SCHEIN PHOTOGRAPHY/CORBIS

Biodiversity: involvement of local people is crucial

SIR — It is true that we can divide the environment up into services that can be priced as discussed in your News Feature “Dollars and sense” (*Nature* 437, 614–616; 2005), even the most sacred. However, there is little evidence that economically sound management systems in themselves engender sustainable utilization. We are frequently reminded of this by bulletins on the state of the world’s native forests or by the collapse in several UK offshore fisheries.

Economically driven management systems often seem to lack the qualities of care and close monitoring that are needed in looking after the rare, vulnerable and endemic. How we manage our environment may be a more fruitful question to ask than why. Revenue from state-run protected areas must be both administered transparently and captured locally: these were rated the most important lessons for sustainability at a 1994 conference on development of protected area strategies for African, Caribbean and Pacific countries.

Similarly, local ‘ownership’ of natural resources and disincentives to cheat may be prerequisites for their long-term sustainable use, as the exemplary state of the Falkland Island fisheries reminds us.

Good interpretation of scientific values and respect for our living heritage also encourage local investment (in the larger sense of the word) in the environment. They are surely as important as economic returns in promoting long-term sustainable use of ecosystem services and the protection of biodiversity.

Martyn G. Murray*, **Jacqueline A. Yelland†**

*Institute of Evolutionary Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK

†2 Wheatsheaf Cottages, Sarratt, Rickmansworth WD3 6BR, UK

Space triumph reveals new spirit of openness in China

SIR — Although I read with interest your news story “China launches plans for space exploration as taikonauts touch down” (*Nature* 437, 1075; 2005), you do not mention a great change that is taking place in China with regards to space missions — a prompt and thorough openness to the public.

Unlike China’s last manned space flight in 2003, the recent mission included a live television broadcast from launch to landing that was made widely available to the public. You could watch images from the electronic cameras installed on the interior and exterior of the module in real time.

Journalists, including those from Taiwan, were also given free entry to the Jiuquan

satellite launch centre, the Beijing aerospace command and control centre, and the landing site in Inner Mongolia. Some journalists were even invited to take part in astronaut training, and try out space suits and food; recording the details of such secret places and instruments would have been inconceivable not so long ago.

More officials have been interviewed by the mass media and openly discussed details of Chinese space exploration technology and plans than ever before. This openness represents a new confidence and self-assurance among scientists in China.

Bin Wang

Department of Pharmacology, Nanjing Medical University, 140 HanZhong Road, Nanjing 210029, P.R. China

Intelligent, social rat can find joy in a hostile world

SIR — One has to admire the guile and tenacity of the lone rat described by James Russell and colleagues (“Intercepting the first rat ashore”, *Nature* 437, 1107; 2005). This rat lived for ten weeks on a small, booby-trapped island, visited by trained rat-killing dogs, before swimming a quarter of a mile to a neighbouring island, where he survived a further two months of concerted efforts to eliminate him.

When the story aired on national public radio in the United States, the tone was contrived to foster the ‘sinister rat’ stereotype. I, for one, sympathize with this castaway doing his best in a hostile world. Russell and his colleagues show us that rats are intelligent, and other studies show them to be social creatures capable of a range of emotions, including joy (see, for example, J. Panksepp and J. Burgdorf *Physiol. Behav.* 79, 533–547; 2003). They deserve more consideration than we give them.

Jonathan Balcombe

Physicians Committee for Responsible Medicine, 5100 Wisconsin Avenue, NW, Suite 400, Washington DC 20016, USA

Peer-review system could gain from author feedback

SIR — The ever-growing number of submissions to many journals has necessarily increased the number of scientists serving as reviewers. Although the majority of these perform their duty honourably and provide valuable feedback to the authors, some produce bad or even damaging reviews, which may not be filtered by the editors.

I believe anonymity is important for the peer-review process, but some power could also be granted to the authors in order to

balance the equation. The flexibility of online systems could be employed to establish a feedback mechanism that may help journals weed out rogue reviewers.

One can imagine a scenario in which all authors would be asked to complete an online questionnaire about the reviews of their manuscript. The questionnaires could be anonymous, but should allow the journal to cross-reference the feedback with the name of each reviewer. Once sufficient data have accumulated, the journal will be able to identify reviewers who are serial offenders and decide not to approach them again.

Gathering feedback from the authors and using that to improve the peer-review process is a simple way of humanizing an increasingly electronic process.

Alon Korngreen

Faculty of Life Sciences and the Leslie and Susan Gonda Brain Research Center, Bar-Ilan University, Ramat-Gan 52900, Israel

How Wallace and Dampier faced tsunamis at sea

SIR — A favourite bedtime book of mine is Alfred Russel Wallace’s *The Malay Archipelago* (Macmillan, London, 1869). During my last read I found the following description of what sounds like a tsunami. Wallace in turn quotes from William Dampier’s *Continuation of a Voyage to New Holland in the Year 1699* (Knapton, London, 1703).

“We heard a dull roaring sound like a heavy surf, ... the roar increased, and we saw a white line of foam coming on, ... as our boat rose easily over the wave. At short intervals ten or a dozen others overtook us with great rapidity, and then the sea became perfectly smooth as before. I concluded at once that these must be earthquake waves; and on reference to the old voyagers we find that these seas have long been subject to similar phenomena.

“Dampier encountered them near Mysol and New Guinea, and described them as: ‘... strange tides, that ran in streams, making a great sea, and roaring so loud that we could hear them before they came within a mile of us. ... These ripples commonly lasted ten or twelve minutes, and then the sea became as still and smooth as a millpond. ... We had one night several of these tides, ... we heard them a long time before they came.’ ...

“Some time afterwards I learnt that an earthquake had been felt on the coast of Gilolo the very day we had encountered these curious waves.”

Jeyaraney Kathirithamby

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK and St Hugh’s College, St Margaret’s Road, Oxford OX2 6LE, UK

COMMENTARY

Climate proofing the Netherlands

Regional climate change should not be seen only as a threat; changes to weather patterns could generate opportunities for large-scale innovations, say **Pavel Kabat**, **Pier Vellinga** and their colleagues.

Throughout human history, people in regions all over the world have learnt, mostly by trial and error, to cope with extreme-climate events. Based on limited climatic and hydrological data from past events, countries have developed infrastructure and legislation to protect people from floods and droughts. Protective measures differ widely between regions, countries and continents — as do the risks. In economically important and densely populated parts of the Netherlands the standards of flood defence are the highest in the world; dykes protect delta regions from a flood event expected to occur once every 10,000 years.

But global climate change caused by greenhouse-gas emissions means that key climate and hydrological variables will change. We can no longer assume that the future climate can be predicted on the basis of past patterns¹. Climate change and sea-level rise present major challenges to each of the world's delta regions, which together harbour about 70% of the world's population and economic

resources. We think the international science and policy communities should develop plans for achieving future sustainability in these vital areas of our planet, using a 'climate proofing' approach.

Climate proofing does not mean reducing climate-based risks to zero — an unrealistic goal for any country. The idea is to use hard infrastructure to reduce risks to a quantified level, accepted by the society or economy. This risk can be further combated by 'softer' measures, such as insurance schemes or, as a last resort, evacuation plans. Such climate proofing should be driven by opportunities for technological, institutional and societal innovations, rather than purely by fear of the negative effects of climate change.

Too little too late

The high impacts of the recent US hurricanes (economic losses associated with Hurricane Katrina were in excess of \$125 billion²), exposed the consequences of not taking enough precautionary measures to address

low-probability but high-magnitude climate events. Most levees in New Orleans, breached by storm waters following Hurricane Katrina, were built to deal with floods that occur once every 30 years. Since Hurricane Katrina hit New Orleans last summer, many have advocated increasing levee protections for New Orleans and even for the entire Louisiana coast. However, a broader climate-proofing approach may be a better long-term solution than simply reinforcing and raising the levees.

Globally, evidence is mounting for more frequent and intense climate extremes in the future, as a consequence of anthropogenic climate change^{3,4}. But these predictions, and those for specific regional impacts, remain uncertain, and deciding on the right strategy to prepare for these events is not an easy task. In the Netherlands, the government is already investing in climate proofing. In addition to the ongoing Climate Changes Spatial Planning Research programme (KvR) scheduled for 2005–2009, which is costing €100 million (US\$118 million), the government will soon launch a new initiative called ARK (Adaptation Programme for Spatial Planning and Climate). ARK will be several times larger than KvR, in both size and scope. It will develop, through partnership between policy makers, researchers and other stakeholders, a comprehensive agenda that deals with climate change across several sectors of the society and economy.

Lost at sea

In the Netherlands, many key decisions about future developments are being taken now, and incorporating climate-change risks and opportunities into these decisions, as was recently called for by the senate of the Dutch parliament, is essential. For the Dutch government, climate change is now accepted as an issue to address in many sectors and policies. But why is it politically acceptable to spend millions of euros on climate proofing in the Netherlands, but not in most other countries?

Sixty percent of the Netherlands territory is located below sea level and 70% of the gross national product is earned in these flood-prone areas. So it is quite likely that the Netherlands will be confronted with several effects of climate change, including increased risk of flooding and more frequent summer droughts. The predictions for the Netherlands'



Vision of the future: a hydrometropole

Today the Netherlands is divided into several dyke-protected regions which have different flooding risks. A future floating city, or hydrometropole, could be further divided so that different risk thresholds are matched to suitable property insurance levels. Finding extra land to store surplus floodwater will require creative solutions.

For example, greenhouse horticulture businesses place a high demand on water for

irrigating their crops, and are sensitive to both wet and dry climate extremes. Greenhouses and their water reservoirs also cover large surface areas. So integrating water reservoirs into the foundations of greenhouses could both save space and serve as emergency floodwater storage⁹. These ideas are already moving from research ideas to pilot projects in the Dutch city of Naaldwijk (see above).

climate in 2100 shown in the table⁵ cover the range of known modelling and emission uncertainties. For most sectors of the Dutch economy and society, even the low to medium climate-change impacts would have serious consequences, such as a significant rise in sea level⁶.

The Netherlands faces several climate-change tests. The first relates to how we cope with increased risk of flooding in regions that are already vulnerable. The second relates to the amount of time we have to adapt to climate change; when does acting later become too late? In our view, developing a climate-proofing strategy now is likely to be more cost effective than taking drastic actions later. The third challenge applies to the way different strategies are being discussed, worked out and eventually implemented and governed. Climate problems call for true integration across scientific disciplines, economic sectors and stakeholder groups. But they also call for a participatory approach in which strategies are discussed at all administrative levels, individual citizens included.

Water people

Public debate about how to cope with future floods, should climate change lead to a greater risk of flooding, was accelerated by two events in the 1990s. Because the Netherlands is situated on the delta of three major European rivers (the Rhine, the Meuse, and the Scheldt), it is especially vulnerable to increased peak discharges. In the winters of 1993 and 1995, extreme discharge nearly overtopped the river dykes, and 250,000 people were evacuated as a precautionary measure.

During the past 50 years, the Netherlands has adopted highly visible policies and measures for water management. After major sea floods in 1953, with economic losses estimated at €1 billion and 2,000 lives lost, the Delta plan was designed. This is a comprehensive system of protective dykes and surge barriers. So far, the Netherlands has invested €13 billion in the Delta plan. Between 1986 and 2005, the main storm surge barrier of the Delta plan has had to be closed more than 20 times because of the threat of flooding.

However, public debate after the 1993 and 1995 floods focused on the potentially negative consequences to the landscape of further raising the dykes. A new 'living with water' strategy was developed⁷, which argues that extreme-climate events should be accommodated rather than fought with heavy infrastructure. The idea is that instead of always reinforcing and heightening the dykes along rivers, occasional flooding will be accommodated and carefully managed in specific designated areas. This new approach, designed to deal with the medium-impact regional predictions for the Netherlands (see Table), was adopted by the Dutch

Regional climate predictions for the Netherlands for 2100 (ref. 5)

Climate variable	Low-impact predictions	Mid-impact predictions	High-impact predictions
Temperature	+1 °C	+2 °C	+4 to 6 °C
Average summer precipitation	+1 %	+2 %	+4 %
Average winter precipitation	+6 %	+12 %	+25 %
Sea-level rise	+20 cm	+60 cm	+110 cm

government as national policy in 2000.

The Netherlands is also starting to build up resilience to other climate changes, such as increased frequency of summer droughts. The summer of 2003 was the hottest in Europe in more than 500 years. In the Netherlands, the heat wave was linked to an estimated 500 deaths, of mostly elderly people, which compared to 27,000 for Europe as a whole. According to some modelling studies⁴, the 2003 summer conditions could become a close-to-normal summer by about 2050. This will have major implications for many sectors of the Dutch economy and society.

For example, the extremely low freshwater discharge by the river Rhine in 2003 resulted in groundwater seepage of seawater to the low-lying delta, which in turn threatened large areas of Dutch agri- and horticulture. New canals for bringing freshwater to the region, and additional summer storage facilities, are now under study. The low 2003 water levels, and high surface-water temperatures, also caused serious shortages of cooling water for the Dutch energy sector, with consequences for industrial productivity.

Paid to protect

What opportunities for climate proofing exist in the agriculture sector? Current changes to the rural policies of the European Union (EU) could be exploited by Dutch farmers wanting to prepare for future climate change. In June 2005, the EU agreed to support the European

Agricultural Fund for Rural Development (2007–2013). This will help reform the Common Agricultural Policy and support

rural development. For many Dutch farmers, this could allow them to diversify their activities and move away from traditional agriculture, such as dairy farming on low-lying peatland.

One possibility is to compensate farmers for lost income. We propose rewarding those farmers achieving lower greenhouse-gas emissions or who use their land as water-storage facilities to combat summer dry spells and winter river discharges. The Dutch dairy farm is currently a significant source of carbon emissions (up to ten carbon dioxide equivalents⁸ per hectare per year). These emission rates could be reduced considerably, and such farms even become a net carbon sink (of up to eight carbon dioxide equivalents⁸ per hectare per year), if the farmland is transformed to a more natural state.

Farming income could be substituted by carbon credits for reduced emissions. The current emissions trading price under the Kyoto Protocol is €23 per ton of carbon dioxide. Assuming that future emissions trading would cover all greenhouse gases and all sectors, which it does not do now, the farmer could generate an income of €414 per hectare per year. For comparison, an average Dutch dairy farm today generates an income of some €670 per hectare⁸ per year, one-third of which is subsidized by the EU.

The Netherlands faces higher sea levels and more extreme hydro-climatic events in the future. We think two basic approaches to climate proofing could help combat these threats. In one, urban and industrial activities, including infrastructure, move from below sea level to higher and drier lands, as found in the eastern Netherlands. The second approach involves the creation of a large 'hydrometropole', a world in which we have learned how to live with — and make a living from — water (see 'Vision of the future: a hydrometropole'). This would be a major urban, industrial and rural area with more than 15 million people living and working in a world partly floating on and surrounded by water. Given the history of the Netherlands and the spirit of its people, this second vision seems more appropriate and attractive, but only time — and vigorous public debate — will tell what approach is favoured.

Pavel Kabat, Wim van Vierssen and Jeroen Veraart are at Wageningen University and Research Centre, Environmental Sciences Group, P.O. Box 47, 6700 AA Wageningen, The Netherlands; Pier Vellinga and Jeroen Aerts are at the Faculty of Earth Sciences, Free University, Amsterdam, The Netherlands. See News Feature, page 275.

"Developing a climate-proofing strategy now is likely to be more cost effective than taking drastic actions later."

1. Kabat, P. & van Schaik, H. *Synthesis Report of the International Dialogue on Water and Climate* 2–25 (ISBN 9032703218, 2003).
2. www.house.gov/transportation/water/10-20-05/10-20-05memo.html#PURPOSE
3. Milly, P. C. D., Wetherald, R. T., Dunne, K. A. & Delworth, T. L. *Nature* **415**, 514–517 (2002).
4. Scott, P. A., Stone, D. A. & Allen, M. R. *Nature* **432**, 610–613 (2004).
5. Können, G. P. *Climate Scenarios for Impact studies in the Netherlands* www.knmi.nl/onderzoek/klimscen/scenarios/Scenarios2001.pdf (2001).
6. Bouwer, L. M. & Vellinga, P. in *Flooding in Europe: Challenges and Developments in Flood Risk Management* (eds Begum, S., Hall, J. W., Stive, M. J. F.) (Springer, in the press).
7. Tielrooy, F. (ed.) *Water Management for the 21st century* (the Hague, the Netherlands, 2000).
8. van den Born, G. J. et al. *Climate Profit in Dutch Fen Meadow Areas: Assessment of Integrated Management Scenarios R-02/05* (Free Univ. Amsterdam, the Netherlands, 2002).
9. ESPACE Guiding Models for Water Storage www.espace-project.org/reports.htm (2004).

COMMENTARY

Policy needs robust climate science

The path between climate science and policy is not always linear, argue **Aristides Patrinos** and **Anjali Bamzai**.

Changes in global and regional climates have always played an important role in civilizations. Migrations from Africa to other continents occurred only when the climate was mild enough to permit them, and cautionary tales are told of entire civilizations collapsing after prolonged periods of drought. The recent catastrophic hurricanes in the United States should be a wake-up call to those concerned with minimizing the impact of such disasters. Societies need robust infrastructures — buildings constructed to withstand high-speed winds, reinforced levees and early warning systems — to deal with extreme weather conditions. Such measures will rely on scientific understanding and accurate predictions of regional climate change, whether it be a result of natural variability or change caused by greenhouse-gas emissions.

In the United States, the Climate Change Science Program (CCSP)¹ is responsible for providing stakeholders and policy makers with the scientific knowledge they need to manage the risks and opportunities of climate change. The CCSP is calling for the delivery of 21 reports to explicitly address the needs of decision makers in sectors such as energy and transport. One such report will focus on understanding regional climate extremes through improved observations and modelling. During 14–16 November, the CCSP held a public workshop on climate science (www.climatechange.gov/workshop2005) to facilitate interactions between researchers and those who rely on the CCSP's products.

Not straightforward

Two papers^{2,3} in this issue highlight the potential impacts of regional climate changes caused by anthropogenic greenhouse-gas emissions. In one Barnett *et al.*² address the effects of regional climate changes on water availability, including in the western United States. The findings reported in both papers are intended to inform policy and decision making. However, there is often controversy about the mechanisms by which scientific research and development (R&D) should do this, especially when it comes to environmental issues.

One model of the interplay between scientific research and policy making describes an almost linear path whereby an environmental problem is identified and R&D is used to investigate the causes, effects and potential

solutions. Information is then fed back to policy makers who may make changes to legislation or technology.

There are instances where problems have been tackled by this linear route; for example, when the discovery of stratospheric ozone depletion led to the Montreal Protocol, which banned substances that harm the ozone layer. But the usual mechanism is less well defined, with repeated interactions occurring between R&D and policy making. We believe that this is the case for regional climate variability. Below we give two examples of multi-level R&D that should eventually give robust solutions for regional climate problems.

"Extreme poverty and limited development in Africa has devastating consequences when combined with natural disasters."

The first describes federal R&D priorities for water availability and usage. The ability to measure, monitor and forecast the state of US and global fresh-water supplies is a problem of national importance. Several federal agencies are developing a research strategy⁴ to understand the processes that control water availability and quality, and to collect baseline information and develop monitoring systems needed to ensure adequate future supplies. Untapped water 'resources' could come from water conservation, water re-use, desalination and aquifer storage. These are generally not considered a resource, but should be.

In addition, recent R&D policy directives⁵ to federal agencies include the development of pilot integrated observing systems for natural-hazards assessment and disaster warnings. Federal agencies will continue to provide strong US leadership for the Global Earth Observing System of Systems (GEOSS)⁶. The US contribution to GEOSS is the planned Integrated Earth Observations System, which is aimed at reducing loss of life and property caused by disasters, and protecting and monitoring our ocean resources.

Our second example of multi-level R&D is the crucial steps taken in response to recent widespread droughts in the western United States. The droughts led governors in these states to unanimously endorse a drought early-

warning system⁷. This system will provide water users across the board — farmers, tribes, business owners, wildlife managers and decision makers at all levels of government — with information that will help them to assess risks in real time, and allow informed decisions to be made ahead of a drought. A national drought policy is being called for that would build upon and enlarge this system.

Vulnerable populations

When developing policy responses to regional climate change, it is essential to consider the vulnerability of the communities affected. For example, extreme poverty and limited development in Africa, coupled with the fragility of ecosystems there, has devastating consequences when combined with natural disasters. The Sahel experienced severe droughts during the 1970s and 1980s, leading to widespread famine and the loss of more than one million human lives. But, paradoxically, excessive development can itself lead to increased vulnerability. Over the past 30 years, US coastal development has quadrupled, and more than 45 million people are now permanent residents of hurricane-prone coastlines. Precarious building in such regions leads to rising economic and human costs associated with natural disasters, as seen with hurricanes Katrina and Rita.

We stand at a crossroads where the risks to humans and property can be minimized by developing robust infrastructures, reliable early warning forecasts and effective response strategies. Maintaining a basic research programme will be critical for success, and the interplay between science and policy will be crucial to how society handles the conflicting dynamics of development and environmental protection. ■

Aristides Patrinos and Anjali Bamzai are at the Office of Science, US Department of Energy, Washington, DC 20585, USA.

See Editorial, page 257; News Feature, page 275.

1. www.climatechange.gov

2. Barnett, T. P., Adam, J. C. & Lettenmaier, D. P. *Nature* **438**, 303–309 (2005).

3. Patz, J. A., Campbell-Lendrum, D., Holloway, T. & Foley, J. A. *Nature* **438**, 310–317 (2005).

4. www.ostp.gov/NSTC/html/swaqreport_2-1-05.pdf

5. www.ostp.gov/html/budget/2007/ostp_omb_guidancememo_FY07.pdf

6. www.epa.gov/geoss/index.html

7. www.westgov.org/wga/initiatives/drought/index.htm

BOOKS & ARTS

The origins of darwinism

Impending anniversaries and the trial over 'intelligent design' make this a good time to revisit Darwin.

From So Simple a Beginning: The Four Great Books of Charles Darwin

edited by Edward O. Wilson

W. W. Norton: 2005. 1,504 pp. \$49.95

Darwin: The Indelible Stamp

edited by James D. Watson

Running Press: 2005. 1,260 pp. \$29.95, £19.99

Bruce H. Weber

As we approach 2009, the 200th anniversary of Charles Darwin's birth and the 150th anniversary of the publication of *On the Origin of Species*, we can expect the current flood of books on Darwin, darwiniana and evolution to become a deluge. So it is hardly surprising that more than one publisher has hit on the idea of producing commemorative volumes from Darwin's *oeuvre*. But even now, the trial continues in a Pennsylvania court over the attempt by the Dover school board to use 'intelligent design' to circumscribe the teaching of evolution.

From So Simple a Beginning and *Darwin: The Indelible Stamp* are both massive editions of Darwin's four major works on evolution. The editors, E. O. Wilson and James Watson, respectively, are distinguished scientists whose contributions to twentieth-century biology have been outstanding and have put them in the public eye.

The same four books are reprinted in each: *The Voyage of the Beagle*; *On the Origin of Species*; *The Descent of Man, and Selection in Relation to Sex*; and *The Expression of the Emotions in Man and Animals*. Both editions provide a general introduction followed by a brief introductory essay for each text. Wilson provides an afterword on evolution and religion that clearly positions the volume in the current argument between evolution and intelligent design.

Both books have iconic and reference value and would sit well on a reader's bookshelf, perhaps next to the complete works of Shakespeare. They are hardly the kind of book that a reader might throw into a backpack to read during the sort of exploration of nature that Darwin so loved. For that purpose there are paperback editions of all these works readily available that provide much more historical contextualization and helpful scholarly support than either of the volumes under review.

For example, Darwin's *The Voyage of the*

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Charles Darwin is still relevant today.

Beagle, edited by Janet Browne and Michael Neve (Penguin, 1989), provides an extensive historical introduction and a biographical guide to the dramatis personae. Ernst Mayr's facsimile of the first edition of *On the Origin of Species* (Harvard University Press, 1964) provides an introduction from his perspective as a founder of the modern evolutionary synthesis. The recent edition of *The Descent of Man* edited by James Moore and Adrian Desmond (Penguin, 2004) provides a ground-breaking historical introduction to this most difficult and least read of Darwin's works and a useful biographical register. For *The Expression of the Emotions in Man and Animals* there is a version edited by Paul Ekman (Oxford University Press, 2002) that contains extensive introductory material, an afterword, appendices and an essay by Philip Prodger about Darwin's pioneering use of photographs (heliotype) in the book.

When approaching Darwin's original works there are two important issues to address. It is imperative to understand what Darwin said, and did not say, in the context of the science and culture of his day. Otherwise we run the risk of reading our present knowledge and

concerns into his words. So much has happened and continues to change in the growth of our scientific understanding of common descent and the theory of natural selection that we need to make connections between Darwin's darwinism and the modern theory of darwinian evolution. Both Wilson and Watson try to provide some historical context and connection to current science, as well as some personal observations about how Darwin's writings affected their intellectual development, but there is too little space devoted to such contextualization. And Watson's volume does not even have an index. Wilson reproduces Darwin's indexes and has generated a unified index for all four texts, keyed to contemporary biological terms and concepts. The reproductions of figures and photographs — paramount for *The Expression of Emotions in Man and Animals* — have been redone for Wilson's edition, and are superior to those in Watson's.

Darwin's works have often been reprinted. Particularly memorable are The Thinker's Library editions, published by the Rationalist Press Association in the wake of the fundamentalist attacks on evolution in the 1920s. C. S. Lewis complained that these cheap scientific books, which sold in the hundreds of thousands despite the economic distress of the times, promoted atheistic materialism. But one also recalls the editions of the Church Fathers that appeared in Darwin's day, produced by different factions interested in co-opting the authority of the past.

The two new reprints are not cheaply done, with the Wilson book having the superior production quality. Given the current controversy in the United States, where polls indicate that only about 10% of the population fully accept a darwinian account of human origins, it is likely that these volumes will play a role in canonizing Darwin's writing. The best outcome will be if Wilson and Watson increase the readership of Darwin, and of the literature on darwinism and evolution. ■

Bruce H. Weber is in the Department of Chemistry and Biochemistry, California State University, Fullerton, California 92834, and in the Division of Science, Bennington College, Bennington, Vermont 05201, USA.

Dancing to Darwin's tune

The Singing Neanderthals: The Origins of Music, Language, Mind and Body
by Steve Mithen

Weidenfeld & Nicolson: 2005. 272 pp. £20

W. Tecumseh Fitch

The years after the publication in 1859 of Darwin's *On the Origin of Species* were full of turmoil. Darwin found a powerfully placed and unyielding opponent in the Oxford linguist Max Müller. The battle lines were clearly drawn: Müller (like many contemporaries) was willing to grant natural selection a role in generating animal form, but he thought it incapable of explaining human evolution, particularly the quintessentially human trait of language. Müller's core critique concerned the lack of plausible precursors of language. He ruthlessly criticized Victorian ideas about the origin of language, coining derogatory nicknames for them, such as the 'bow wow' and 'ding dong' theories, which are still used today.

Darwin had cautiously omitted virtually all mention of human evolution in *On the Origin of Species*, but he took on the problem of language evolution in his 1871 book *The Descent of Man, and Selection in Relation to Sex*. Like most subsequent theorists, Darwin recognized that something as complex as human language could not have sprung into existence fully formed, but must have passed through some intermediate phase (today termed a 'protolanguage'). Recognizing that music is a human universal, rendered peculiar by its lack of any obvious survival function, he elaborated on an idea briefly mentioned by Jean-Jacques Rousseau that protolanguage was similar to music, and suggested that "the rhythms and cadences of oratory are derived from previously developed musical powers". He dismissed the converse idea, put forward by Denis Diderot and Herbert Spencer, that music derives from speech, as being contrary to basic evolutionary principles. Darwin argued, by analogy with birdsong, that early musical capacities were driven by sexual selection: "Musical notes and rhythm were first acquired by the male or female progenitors of mankind for the sake of charming the opposite sex."

Despite recent scholarly interest in the evolution of language, *The Singing Neanderthals* by Steven Mithen is the first book-length exposition of Darwin's 'musical protolanguage' hypothesis that I have seen. Mithen, a professed non-musician, mounts an enthusiastic

and often impressive defence of Darwin's hypothesis, covering a wide variety of fascinating topics.

Mithen first takes aim at Spencer's idea that music is simply a non-adaptive by-product of spoken language — a viewpoint memorably expressed by Steven Pinker in his 'music as cheesecake' hypothesis in *How the Mind Works* (W. W. Norton, 1997). Mithen argues that the power of music over human emotions and behaviour is inconsistent with this hypothesis. He ably summarizes current proposals for adaptive functions of music, covering its roles in childcare, group cohesion, competition and



Sound idea: did music develop through parents entertaining their children?

mate choice. A childcare function seems the most convincing, as parents in all human cultures sing to their children, and experimental work has shown that mothers' lullabies and play songs have an important role in regulating infant arousal. But each of the other alternatives receives some support from Mithen's review. He has assembled a considerable body of data and argument that modern-day Müllers, critical of Darwin's proposal of a musical protolanguage, will need to answer.

Mithen then builds on his expertise in human palaeontology and archaeology to develop a detailed picture of this hypothetical musical protolanguage, which he dubs "Hmmm", for "holistic, manipulative, multimodal, musical and mimetic". Although Mithen's scenario is far more detailed than previous accounts, I was surprised and disappointed that nowhere in this long volume, rich with footnotes, does he acknowledge that the core idea of the book is Darwin's. He cites *The Descent of Man*, and some relevant passages are even quoted in passing, but Mithen seems intent on claiming the musical protolanguage hypothesis as his own. The uninformed reader might easily finish the book without realizing that the central thesis of the book was concisely

laid out by Darwin more than a century ago. "Hmmm", indeed.

The book's biggest weakness is that Mithen's enthusiasm for the musical protolanguage hypothesis sometimes prevents him from frankly acknowledging its problems. This is most clear in his treatment of the neuroscience of music. He reviews brain-lesion studies that indicate a separation of musical and linguistic circuitry in the brain, providing evidence against the idea that music is simply a by-product of language. But he fails to review more recent brain-imaging work that paints a more complex picture of partly overlapping neural circuitry for music and language. His discussion of potential fossil cues to the evolution of song and speech is outdated and incorrect in several places, and his treatment

of potential parallels between music and animal vocalizations is almost entirely limited to primates, ignoring a wider range of animal data, such as birdsong, that is relevant to the evolution of music. So although the book provides broad coverage of supporting data, readers seeking a balanced overview will have to turn to the specialist literature.

These criticisms aside, the book is extremely well-written, and Mithen's clear and infectious enthusiasm make it a good introduction for non-specialists interested in the topic. I can recommend it to anyone interested in the biology and evolution of music or

language — and particularly to readers interested in Darwin's idea that music constitutes an ancient and important form of human communication, intertwined with, but independent from, language.

W. Tecumseh Fitch is in the School of Psychology, University of St Andrews, St Andrews, Fife KY16 9JU, UK.

DARWIN IN FICTION

Darwin's life and work has also inspired two recent novels.

In *This Thing of Darkness* (Review, £17.99), Harry Thompson focuses on the relationship between Charles Darwin and Robert Fitzroy, captain of *The Beagle*, and the origins of Darwin's theory of evolution and the religious debate it sparked. This fictionalized biography was long-listed for the 2005 Man Booker Prize.

John Darnton's *The Darwin Conspiracy* (Knopf, \$24.95) is a novel with a darker secret, told through three intertwined narratives. Two modern-day scholars trace Darwin's research path and also uncover the diaries of Darwin's daughter Lizzie.



BAYERISCHE STAATSGEMÄLDESAMMLUNGEN, NEUE PINAKOTHEK, MUNICH

Monkey business

Charles Darwin changed the way animals were viewed in art.

Colin Martin

Between 1750 and 1900, at a time of accelerated scientific advance and industrialization in Europe, there was a marked change in the human perception and appreciation of animals. This shift can be seen in the paintings, drawings, prints and sculptures exhibited in *Fierce Friends*, an exhibition that can be seen at the Van Gogh Museum in Amsterdam, The Netherlands, until 5 February 2006. The curators' careful selection of artworks and other exhibits demonstrates how our view of other animals changed as we developed different ways of looking at the world.

In the eighteenth century there was a rapid growth in the number of animals known to Europeans, as new species were discovered and brought from overseas. In the early nineteenth century, the use of dissection increased our knowledge of their comparative anatomy and physiology. In the mid-nineteenth century, geology and the

new science of palaeontology unearthed fossil data that challenged religious beliefs about creation. And in the late nineteenth century, Darwin's theory of evolution helped us understand that new species evolved from older ones.

It was the publication of Darwin's *On The Origin of Species* in 1859 that brought the connection between the human and animal worlds into sharpest focus and had the most profound effect on our perception of animals. *Homo sapiens* could no longer be considered to be 'above' other animal species, but was inextricably linked to them, most closely to primates.

Fierce Friends includes examples from the tradition of painting monkeys dressed like humans and engaged in human occupations, from Jean-Baptiste-Siméon Chardin (in around 1740) and Alexandre Decamps (1837). In contrast, Gabriel Max's 1889 painting *The Jury of Apes* reflects a post-darwinian approach to depicting monkeys. Max

reportedly painted the picture to express his unhappiness about the works selected for the first annual art exhibition in Munich. But he uses monkeys to criticize the jury in a new way. He was a darwinist who corresponded with German zoologist Ernst Haeckel, a promoter of evolutionary theory.

Max kept monkeys as pets and his painting shows how carefully he must have observed them. The many different species in his jury are identifiable and behave naturally, despite being shown in an unnatural habitat, clustered together on a packing crate. They are not dressed up or caricatured. If Max intended to criticize the jurors, he does so by subtly implying that the jurors have lost their natural instincts when confronted by art, unlike the monkeys in his painting. The Munich jurors might be stupid, but the monkeys are not.

Fierce Friends can also be seen at the Carnegie Museum of Art in Pittsburgh from 25 March to 28 August 2006. Colin Martin is a London-based writer.

A polymath's dilemma

Thomas Young strove to satisfy his curiosity in virtually every scientific subject and, undeterred by sceptics calling for a narrower focus, made discoveries in almost all the fields he studied.

Andrew Robinson

Polymaths have always posed a problem in academia. How do they relate to specialization and interdisciplinarity, genius and dilettantism, inspiration and perspiration? Robert Hooke, Benjamin Franklin and Alexander von Humboldt were among those who were too academically wide-ranging for posterity to cope with, and their scientific reputations suffered as a consequence.

Individual curiosity is the driving force of science, but when insatiable, can it hamper the intellectual? The life and work of the polymath Thomas Young (1773–1829) illuminates the issue perhaps more acutely than that of any other scientist. Today, views of Young span the spectrum from near-universal genius to dabbling dilettante. Those who appreciate him — especially physicists, physiologists and Egyptologists — admire his range, his intuition and his far-sightedness. Those who do not, depreciate these same aspects of his life and work as sloppiness and opportunism.

Some great names of nineteenth-century science, notably John Herschel, Hermann von Helmholtz and John William Strutt (Rayleigh), were in awe of Young. In 1931, Einstein paid tribute to him in a brief foreword to Newton's *Opticks*; he referred to Newton's observations of the colours of thin films "as the origin of the next great theoretical advance, which had to await, over a hundred years, the coming of Thomas Young." In *Nature*, Joseph Larmor, a former Lucasian professor of mathematics at Cambridge, wrote an essay on Young calling his 1802–03 lectures on natural philosophy at the Royal Institution "the greatest and most original of all general lecture courses". In 1973, on Young's bicentenary, the Science Museum in London noted, startlingly, that "Young probably had a wider range of creative learning than any other Englishman in history. He made discoveries in nearly every field he studied."

Young made a pioneering contribution to the understanding of light by demonstrating interference patterns, known as 'Young's fringes', around 1800, which led to the Young–Fresnel undulatory theory. He also formulated an important measure of elasticity, called 'Young's modulus'. He was the first to explain the accommodation of the eye; he discovered the phenomenon of astigmatism; and he proposed the three-



World explorer: from the human eye to Egyptian script, Thomas Young's interests ranged widely.

colour theory of vision. This was later known as the Young–Helmholtz theory, and was finally confirmed experimentally in 1959. He undertook seminal detective work on the Rosetta Stone and helped to found Egyptology. Although the credit for finally reading the hieroglyphs belongs to Jean-François Champollion, Young was the decipherer of the second type of Egyptian script on the Rosetta Stone, known as demotic script.

In addition, he was a distinguished physician at St George's Hospital; foreign secretary of the Royal Society for a quarter of a century; an authoritative writer on all manner of subjects; a major scholar of ancient Greek; and a phenomenal linguist who coined the term 'Indo-European' for the language family that includes Greek and Sanskrit.

When pressed to contribute to the *Encyclopaedia Britannica*, Young offered articles on the alphabet, annuities, attraction, capillary action, cohesion, colour, dew, Egypt, the eye, focus, friction, haloes, hieroglyphics, hydraulics, motion, resistance, ships, sound, strength, tides, waves and "anything of a medical nature". And he wasn't boasting: having been an 'inspector of calculations' and physician of a London-based life-insurance company in the 1820s, he knew about annuities. And his roles as

adviser to the Admiralty on shipbuilding, secretary of the Board of Longitude, and superintendent of the vital *Nautical Almanac* from 1818 until his death had informed him on ships.

He also wrote many biographical articles about scientists and mathematicians, an occupation that led him to reflect on his own intellectual motivation to a close friend: "The biographical articles seldom amuse me much in writing; there is too little invention to occupy the mind sufficiently: I like a deep and difficult investigation when I happen to have made it easy to myself if not to all others — and there is a spirit of gambling in this, whether as by the cast of a die, a calculation *à perte de vue*, shall bring out a beautiful and simple result, or shall be wholly thrown away."

Scarcely the words of a dilettante. But, on the other hand, Young was restlessly curious. He generally moved on long before he had fully explored his intuitions and discoveries. As a result, his reputation suffered, which he well knew. "Whether the public would have been

more benefited by his confining his exertions within narrower limits, is a question of great doubt," Young said in an autobiographical sketch intended for a posthumous edition of the *Britannica*. After his death, the president of the Royal Society could not help but echo this ambivalence towards polymathy in a valedictory address: "[His] example is only to be followed by those of equal capacity and equal perseverance; and rather recommends the concentration of research within the limits of some defined portion of science, than the endeavour to embrace the whole."

Whether one admires polymaths seems a matter of taste, not objective judgement. But it should surely be indisputable that a man of sweeping vision like Young has a place in science as valuable as, say, the more narrowly focused Augustin Fresnel, Helmholtz or Champollion. In Young's own perceptive words: "It is probably best for mankind that the researches of some investigators should be conceived within a narrow compass, while others pass more rapidly through a more extensive sphere of research."

Andrew Robinson is the author of a biography of Thomas Young, *The Last Man Who Knew Everything*, which is to be published by Pi Press in January. He is literary editor of *The Times Higher Education Supplement*.

BRIDGEMAN.CO.UK

ESSAY

NEWS & VIEWS



K. HESSE/GETTY IMAGES

EPIDEMIOLOGY

Dimensions of superspreading

Alison P. Galvani and Robert M. May

Analyses of contact-tracing data on the spread of infectious disease, combined with mathematical models, show that control measures require better knowledge of variability in individual infectiousness.

The SARS epidemic was notable for the existence of 'superspreaders' who infected dozens of people, whereas other infectious individuals infected few or none. Were SARS superspreaders anomalies, or are superspreaders characteristic of most infectious diseases? What effects does heterogeneity in infectiousness have on disease emergence and control? On page 355 of this issue, Lloyd-Smith *et al.*¹ provide insight into such questions, and more.

The first question any ecologist asks about an invasive species is: what is the invader's intrinsic capacity for population increase? To answer this, the species' basic reproductive number, R_0 , is measured by the average number of offspring per capita that survive to reproductive age. For a directly transmitted infectious disease, be it polio, smallpox, SARS, HIV/AIDS or some newly emerging pathogen, R_0 is the average number of infections produced by an infected individual in a susceptible population². If R_0 is less than one, a self-sustaining epidemic is not possible (at least without further pathogen evolution). If R_0 exceeds one, then although early stochastic fluctuations may extinguish the invader, an epidemic is possible. If R_0 is large, an epidemic is virtually certain.

Initial work in this area largely treated individuals in populations as having an equal chance of transmitting disease — that is, as being homogeneous — and ignored stochastic fluctuations in transmission capability. However, studies of gonorrhoea³, and of HIV/AIDS⁴, could not explain epidemiological patterns without acknowledging heterogeneities in patterns of sexual-partner acquisition, including the disproportionate influence of superspreaders. Similarly, knowledge of heterogeneous parasite burdens is fundamental to accurate modelling of helminthic diseases^{5,6}. Explanations of epidemiological patterns of malaria also depend on understanding heterogeneous biting by the mosquito vector⁷.

These observations led to the proposal of the 20/80 rule^{2,8}, which suggests that roughly 20% of the most infectious individuals are responsible for 80% of the transmission (Fig. 1, overleaf). This rule has been applied mainly to helminthic and sexually transmitted diseases⁷; for other directly transmitted diseases, such as smallpox or influenza, heterogeneity in infectiousness has been neglected. The superspreading that seemed to fuel the 2003 SARS epidemic was largely treated as anomalous in most models, but it highlighted the need for a

reassessment of heterogeneous infectiousness⁹.

Lloyd-Smith *et al.*¹ address this point by posing infectiousness as a continuous variable, and formulate an unambiguous and universally applicable definition of superspreaders as those who transmit more infection than is predicted by a homogeneous 'null model'. The authors analyse data from eight human infections, including SARS, measles, smallpox, monkeypox and pneumonic plague, to show that superspreading occurs across the board, although to a greater or lesser extent depending on the disease. Heterogeneity is greatest for SARS and least for Ebola haemorrhagic fever.

Analysis of the epidemiological dynamics shows that, for a given R_0 , both the probability that an epidemic will take off, and the subsequent course of the epidemic, are affected by such heterogeneity. These results may be appreciated intuitively. For a given value of R_0 , high heterogeneity in infectiousness implies that relatively few individuals are responsible for most of the transmission — or conversely, that many individuals do not transmit at all. In turn, such small numbers tend to generate pronounced stochastic fluctuations in the initial stages of the epidemic. Consequently, a heterogeneously infectious emerging disease will be

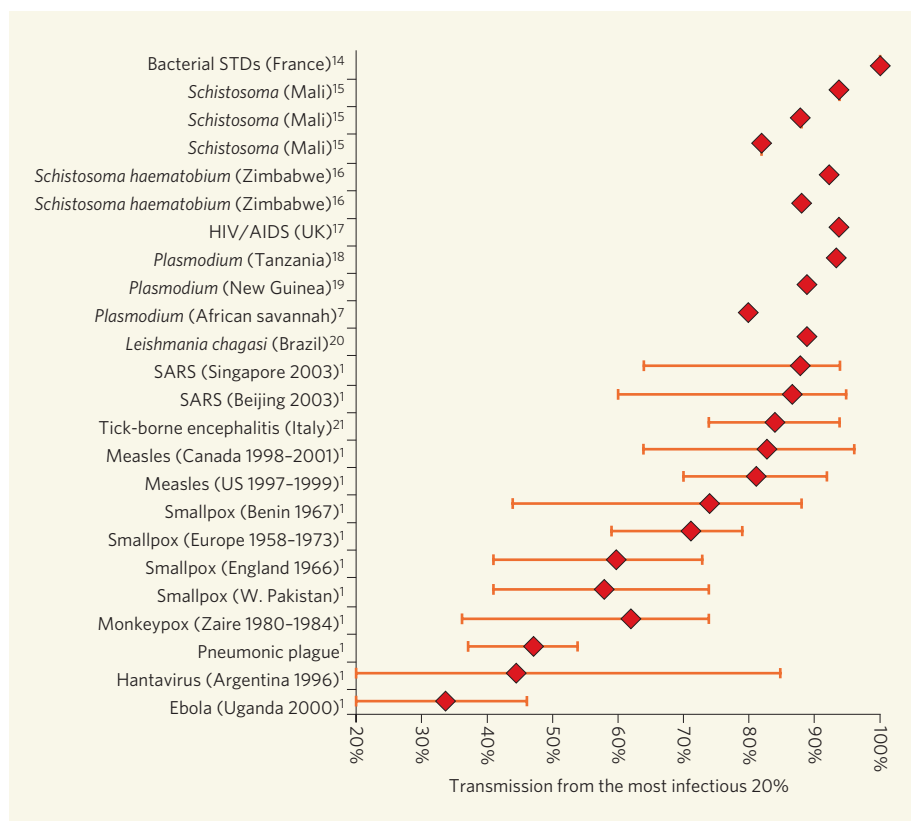


Figure 1 | Heterogeneity in infectiousness for a range of diseases. The measure used is the 20/XX index, which quantifies the proportion of the transmission (XX%) that results from the most infectious 20% of the population. Confidence intervals are included where available. The interval for tick-borne encephalitis (Italy)²¹ indicates possible values depending on assumptions made about host susceptibility. STDs, sexually transmitted diseases; HIV, human immunodeficiency virus; AIDS, acquired immunodeficiency syndrome; SARS, severe acute respiratory syndrome.

less likely to generate an epidemic, but if sustained, the resulting epidemic is more likely to be explosive. Thus, it is dangerous to underestimate a disease on the basis of frequent 'failed' attempts, as exemplified by bird flu.

The authors highlight the practical implications of their work. Control efforts should aim to identify the highly infectious super-spreaders, and target vaccination or other interventions at them. In this way, the outbreak may be halted sooner, and with fewer people treated, than if efforts are directed at random individuals. Furthermore, Lloyd-Smith *et al.* distinguish between individual-specific and population-wide control measures (for example isolating individual patients as opposed to advising an entire population to reduce the behaviours associated with transmission). They show that individual-specific strategies are more likely to exterminate an emerging disease than population-wide interventions, because the former increase heterogeneity in infectiousness.

Comparisons can be drawn between these new findings¹ and work on heterogeneities in contact patterns among individuals (for example in SARS¹⁰, or sexually transmitted diseases such as HIV/AIDS⁴) and among groups (foot-and-mouth disease, for instance¹¹). Earlier research on contact patterns led to a generalized theorem (ref. 2; equation 12.23), which

was based on the following reasoning. We can estimate the proportion, p^* , to be vaccinated or otherwise treated in order to eradicate infection in a homogeneous population ($p^* = 1 - 1/R_0$). However, if we take advantage of the heterogeneity, and target the more infectious or more sexually active individuals (depending on the disease), then we can achieve our aim by treating a smaller proportion than is estimated by p^* , as echoed in the results of Lloyd-Smith and colleagues. This general result, first discovered in the epidemiological literature in the mid-1980s, also applies to the structure of information-technology networks in relation to targeted versus random 'viral' attacks¹².

There are also differences between the work of Lloyd-Smith *et al.* and work on heterogeneities in contact patterns. Given that contact rates govern the likelihood both of becoming infected and of passing on infection, models based on heterogeneous contact rates have assumed perfect correlation between infectiousness and susceptibility. Consider HIV/AIDS, where in the simplest case $R_0 = \beta Dc$, with β being the transmission probability (a measure of the infectiousness of an infected individual), D the duration of the infectiousness, and c the average rate at which new sexual partners are acquired. Heterogeneities among individuals with respect to β

or in D do not directly affect R_0 as such: the quantities β and D enter the dynamic equations linearly, and the appropriate values for estimating R_0 are just the simple averages.

By contrast, the distribution of partner-acquisition rates enters nonlinearly; those with more partners are more likely to acquire infection by virtue of their higher activity, and they are also more likely to transmit infection. Consequently, the epidemiologically appropriate 'average partner-acquisition rate', c , is not the mean of the distribution, but rather the mean-square divided by the mean. An incorrect result is obtained for R_0 when the average of the partner-acquisition or contact distribution is used. (This observation, incidentally, helps explain how large geographical variation in HIV incidence can arise from differences in the tails of such distributions.) In contrast, Lloyd-Smith *et al.* evaluate heterogeneity in overall R_0 (integrating all contributing factors), and assume that infectiousness is not correlated with susceptibility. As they note, the reality probably lies somewhere in between, with some intermediate level of correlation between infectiousness and susceptibility.

Although there is a considerable advantage in targeted control measures if highly infectious individuals can be identified before they have transmitted infection^{1,8}, this is easier said than done. In the case of sexually transmitted diseases and contact patterns, however, Cohen *et al.*¹³ formulated a seemingly paradoxical method for achieving this aim, without directly identifying the active individuals. This procedure is based on the realization that one's contacts will on average be more highly connected within a contact network than oneself, simply by virtue of being a contact. Thus, highly connected individuals can be identified for intervention by first picking individuals at random, and then selecting randomly among their acquaintances. In this way, highly connected individuals are identified with minimal effort. Moreover, this procedure can be carried out either before or after an outbreak.

Among the next steps to be taken are further parametrization of heterogeneity for different diseases and for the same disease in different settings, and determination of the characteristics of emerging diseases that are likely to exhibit the most pronounced heterogeneity. Heterogeneous infectiousness, and its extreme manifestation of superspreading, are likely to be general properties of disease transmission in populations. The ambitious aim of controlling disease emergence will require a better understanding of those properties, which is most likely to be achieved through the combination of data analysis and epidemiological theory exemplified by Lloyd-Smith and colleagues' study. ■

Alison P. Galvani is in the Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut 06520, USA.

Robert M. May is in the Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. e-mails: alison.galvani@yale.edu; robert.may@zoo.ox.ac.uk

- Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. *Nature* **438**, 355–359 (2005).
- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, 1991).
- Hethcote, H. W. & Yorke, J. A. *Lect. Notes Biomath.* **56**, 1–105 (1984).
- May, R. M. & Anderson, R. M. *Nature* **326**, 137–142 (1987).
- Grenfell, B. T., Wilson, K., Isham, V. S., Boyd, H. E. G. & Dietz, K. *Parasitology* **111**, S135–S151 (1995).
- Galvani, A. P. *J. Parasitol.* **89**, 232–241 (2003).
- Smith, D. L., Dushoff, J., Snow, R. W. & Hay, S. I. *Nature* (in the press); doi:10.1038/nature04024.
- Woolhouse, M. E. J. *Proc. Natl Acad. Sci. USA* **94**, 338–342 (1997).
- Bauch, C. T., Lloyd-Smith, J. O., Coffee, M. & Galvani, A. P. *Epidemiology* **16**, 791–801 (2005).
- McLean, A. R., May, R. M., Pattison, J. & Weiss, R. A.

- SARS: A Case Study in Emerging Infections* (Oxford Univ. Press, 2005).
- Haydon, D. T. et al. *Proc. Biol. Sci.* **270**, 121–127 (2003).
 - Albert, R., Jeong, H. & Barabási, A. L. *Nature* **406**, 378–382 (2000).
 - Cohen, R., Havlin, S. & Ben-Avraham, D. *Phys. Rev. Lett.* **91**, 247901–247904 (2003).
 - Spira, A. & Bajos, N. *Sexual Behavior and AIDS* (Avebury, Aldershot, 1994).
 - Etard, J. F., Audibert, M. & Dabo, A. *Am. J. Trop. Med. Hyg.* **52**, 549–558 (1995).
 - Chandiwana, S. K. & Woolhouse, M. E. *Parasitology* **103**, 363–370 (1991).
 - Johnson, A. M., Wadsworth, J., Wellings, K. & Field, J. *Sexual Attitudes and Lifestyles* (Blackwell Scientific, Oxford, 1994).
 - Smith, T., Charlwood, J. D., Takken, W., Tanner, M. & Spiegelhalter, D. J. *Acta Trop.* **59**, 1–18 (1995).
 - Hij, J. L. et al. *J. Med. Entomol.* **34**, 193–205 (1997).
 - Quinnell, R. J. & Dye, C. *Med. Vet. Entomol.* **8**, 219–224 (1994).
 - Perkins, S. E., Cattadori, I. M., Tagliapietra, V., Rizzoli, A. P. & Hudson, P. J. *Int. J. Parasitol.* **33**, 909–917 (2003).

NANO-OPTICS

Gold loses its lustre

Roy Sambles

The perfect lens would immaculately reproduce an image of an object, with no light losses in the transition. The strange optical properties of a gold nanostructure bring the prospect of such a component into sharper focus.

As one luckless wooer in Shakespeare's *The Merchant of Venice* discovers, all that glisters is not gold. But what if gold did not 'glisten' at all; what if it could, in fact, be made transparent? Such a material would be precious in itself — a potential basis for a 'perfect' lens. Writing in this issue, Grigorenko and colleagues (page 335)¹ present convincing evidence that they have produced nanostructured gold with remarkable optical properties. Although not quite perfect lens material, what they have made is a significant step towards that end.

Grigorenko and colleagues' gold demonstrates, when illuminated by visible light of certain polarizations and at certain incident angles, a characteristic known as negative permeability. To understand the context of this statement, we require some definitions. First, the permeability, μ , of a material expresses the extent to which an applied magnetic field is enhanced in that material: the higher the permeability, the more magnetic a material can become. A second, similar quantity, the permittivity ϵ of a material, relates to electric fields. In this case the definition is slightly different: large, positive permittivities are found in materials — namely insulators, or 'dielectrics' — that respond to an externally applied electric field to produce a distribution of stored charge that reduces the electric field

within them. Finally, both the permeability and the permittivity of a material are related to its refractive index, n — the degree to which it bends incident electromagnetic radiation, such as light. This relationship is defined by the formula $n = (\mu\epsilon)^{1/2}$.

So why is the negative permeability of Grigorenko and colleagues' material exciting? In answering this, it is important to appreciate

that even normal metals are quite extraordinary in their response to light. Free electrons within metals readily respond to the electric field of incident electromagnetic radiation and thereby cancel it almost completely, provided that this field does not oscillate too quickly; so below a certain field frequency, called the plasma frequency, the real part of a metal's optical permittivity is negative. (Expressing permittivity as a complex number with a real and imaginary part is a mathematical construct that allows the wave nature of the fields involved to be taken into account; the imaginary part of permittivity, scaled by the imaginary unit i , is associated with the scattering of electrons and resultant heating in the material.)

Gold, for an incident electric field at red wavelengths, has an optical permittivity of about $-10+2i$, coupled with a normal, positive permeability. Taking these facts into account and using the formula for n , it can then be calculated that the refractive index of gold must be almost entirely imaginary. This is the mathematical equivalent of saying that the metal is opaque — it acts as a barrier to light, with the amplitude of the incident electric field decaying exponentially once inside the surface.

If the permeability of a metal such as gold were to be negative instead of positive, however, it turns out that it would have a negative refractive index^{2,3}. Such a material will bend light in the opposite direction to normal materials, lending them their potential as perfect lenses³: a flat sheet of the material would focus the light to a perfect image on the other side of the sheet (Fig. 1).

This concept of materials of negative refractive index has been tested in the microwave region of the electromagnetic spectrum⁴. Here, it proved not too difficult to fabricate a resonant metallic material from components known as split-ring resonators, which have both negative permittivity and negative permeability for a small range of incident frequencies. But making a similar material that is responsive at higher frequencies in the visible range is not so easy, as it would require nanoscale split-ring resonators. Grigorenko and colleagues' contribution¹ is to overcome this barrier to a certain extent. They use nanofabrication procedures to make a patterned surface comprising tapered gold posts arranged periodically in pairs. Over a limited frequency range in the visible spectrum, these pairs behave as small, high-frequency bar magnets, much as split-ring resonators do when used at microwave frequencies. A characteristic of such bar magnets at optical frequencies is that they act to cancel the magnetic component of the incident radiation (Fig. 2, overleaf) — much like the action of the electrons in a metal is to

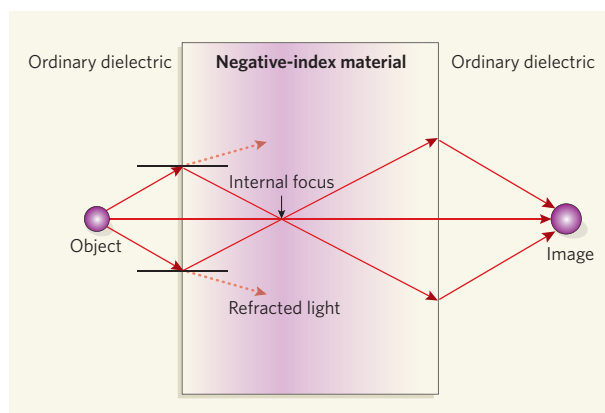


Figure 1 | Reverse swing. Light waves (arrows) from an external source will, at the interface between two materials of different refractive indices, bend towards or away from the normal to the interface (dotted arrows) but never beyond the normal. This limitation is overcome if one of the materials has a negative refractive index. The same thing happens at the second interface of the material, so it acts as a perfect lens, reproducing an image of an object. A conventional lens, which requires a curved surface, can never produce a perfect image because it will always fail to refocus the light that comes from the object in the form of decaying (evanescent) waves. Thus the image will not contain the information about the object carried by these waves.

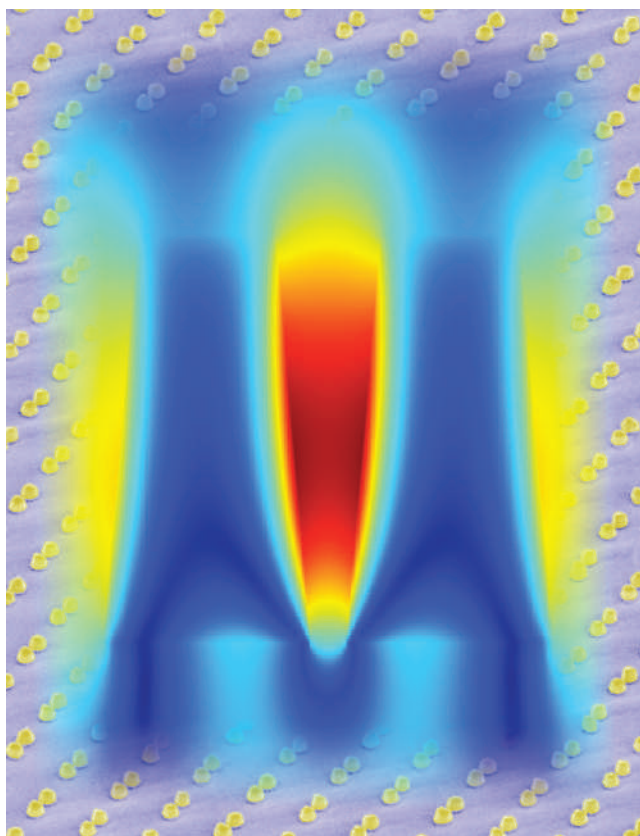


Figure 2 | The negative consequences of magnetism. The distribution of the magnetic field excited by light around a pair of gold nanoposts in Grigorenko and colleagues' study¹. The suppression of the magnetic field by the action of the bar magnets that leads to the structure's negative permeability is shown by the areas coloured in blue. The image is superimposed on a micrograph of an ensemble of nanopost pairs. (Courtesy of A. Grigorenko and colleagues.)

cancel the electric field. This leads to the gold structure having negative permeability.

Were it not for the rather large imaginary contribution to their material's permeability, Grigorenko and colleagues would already have found the way to negative refraction. Although their achievement stops short of this, they were able, by matching the impedance (defined as the ratio ϵ/μ) of their patterned gold to that of an adjacent dielectric, to stop it reflecting. This is in itself a significant step towards a perfect lens, and other novel optical components for visible frequencies.

Further hurdles remain to be overcome. Reducing the imaginary contribution to the optical permeability will be no trivial task. It is also not obvious how structures such as those developed by Grigorenko and colleagues¹ might be made three-dimensional. Nevertheless, it seems that what nanopatterned gold is losing in glister, it is gaining in transparency. ■

Roy Sambles is in the School of Physics, University of Exeter, Stocker Road, Exeter EX4 4QL, UK.

e-mail: j.r.sambles@exeter.ac.uk

1. Grigorenko, A. N. *et al. Nature* **438**, 335–338 (2005).
2. Veselago, V. G. *Sov. Phys. Usp.* **10**, 509–514 (1968).
3. Pendry, J. B. *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
4. Parimi, P. V., Lu, W. T., Vodo, P. & Sridhar, S. *Nature* **426**, 404 (2003).

PALAEONTOLOGY

Data on a plate

The weird fossil pictured here is a specimen of *Ceratocystis* — a member of an ancient group of animals, the Stylophora, that lived roughly 500 million to 300 million years ago. Stylophorans, which were just centimetres in length, bear little resemblance to any extant animal and have defied categorization for decades.

Sébastien Clausen and Andrew B. Smith, in a report elsewhere in this issue (*Nature* **438**, 351–354; 2005), claim to have broken the deadlock. They have identified a single feature which, they argue, rules out two of three hypotheses about the biology of stylophorans. Figure 1 of their paper (page 351) offers a quick guide to the hypotheses under test.

The body of *Ceratocystis* is divided into a large, irregular-plated blob at one end; a long, thin, segmented section at the other (apparently constructed as a column of discs); and a connecting region that looks like bellows run over by a lawnmower. The only feature on which all authorities agree concerns the various plates: they look exactly like

the calcitic plates of echinoderms, a large group of marine organisms that today includes starfishes, sea urchins and the like.

This assignment has given rise to the three hypotheses. The most enduring is that the long segmented section and the connecting region comprise a mobile stem. Many fossil echinoderms, and some extant ones — the crinoids — have similar appendages. According to this view,



stylophorans are very primitive echinoderms that evolved before the appearance of other echinoderm features, such as the distinctive water-vascular system manifested as arrays of canals ending in avenues of 'tube-feet'.

In one of the alternative hypotheses, stylophorans are held to be highly evolved echinoderms in which the stalk is a feeding arm, with a mouth somewhere in the middle linked to tube-feet that are covered by retractable plates. In the other hypothesis, they are interpreted as primitive chordates that retain a

calcite skeleton from a more remote common ancestor of echinoderms and chordates, and the stalk contains muscle blocks, a notochord and a brain.

Clausen and Smith tackle the job of inferring the soft parts of *Ceratocystis* by studying the microstructure of the calcite plates. The mid-stalk region of the creature bears one large ossicle known as the stylocone, and the surface structure of

the calcite (replaced by iron oxides in the specimens studied) has textures that, by analogy with modern echinoderms, give an indication of the kind of tissue to which the stylocone was adjacent in life.

It seems that the part of the stylocone adjoining the long segmented section faced connective tissue, as is now seen in modern stalked echinoderms such as crinoids. In contrast, the part of the stylocone next to the bellows region shows a surface similar to those that in echinoderms act as attachment sites for muscle. There are no signs of anything like a mouth or tube-feet as implied by the feeding-arm model, and the plates in the stalk do not seem to have been hinged to allow exposure of tube-feet. The chordate hypothesis is also ruled out, as the muscle would have inserted directly into the calcite, rather than being bound up in discrete, chordate-style muscle blocks.

The conclusion, then, is that *Ceratocystis* used its appendage as a muscular, locomotory organ. So its anatomy (and its evolutionary position) conforms to the oldest and least demanding of the three hypotheses.

Henry Gee

DEVELOPMENTAL BIOLOGY

The X-inactivation yo-yo

Wolf Reik and Anne C. Ferguson-Smith

In female mammals, one of two X chromosomes has to be shut down during early development. To what extent does this 'imprinted X-chromosome inactivation' involve the history of the chromosome?

In most mammals, males have the male sex-determining Y chromosome and a single X chromosome, whereas females have two X chromosomes. In females, the resulting imbalance in the 'dosage' of genes on the X chromosomes needs to be compensated so that gene expression from the X chromosome is equivalent in males and females. Mammals have evolved a unique form of dosage compensation, called X-chromosome inactivation, in which one of the two X chromosomes in female cells is silenced epigenetically¹ — that is, by factors such as chemical modification of the DNA, or of the histone proteins that package DNA into chromosomes, often involving non-coding RNAs. Many aspects of mammalian X inactivation remain mysterious. But through elegant studies in the mouse, Okamoto and colleagues (page 369 of this issue)² have unravelled some of the earliest events in the process.

During early development of female mouse embryos, and in extra-embryonic tissues such as the placenta, it is always the X chromosome derived from the father that is inactivated³. Gene expression from only one parental member of a chromosome pair is known as imprinting, and is caused by an epigenetic memory first arising in the egg or the sperm. Later on, in the embryonic tissues, X inactivation is random with respect to the parental origin of the X chromosomes⁴.

There is considerable interest in understanding the mechanisms that specifically silence the paternal X chromosome in early development, and the extent to which the history of that chromosome is involved. During male meiosis, in which sperm are produced, a process known as meiotic sex chromosome inactivation (MSCI) occurs. In developing male sperm, the sex chromosomes form a unique structure, the XY body; MSCI occurs here and leads to repression of the transcription of X- (as well as Y-) linked genes⁵. This meiotic inactivation uniquely affects the sex chromosomes and may be associated with the inability of the X and Y chromosomes to pair during male meiosis⁵.

One proposal^{6,7} is that, when an egg is fertilized, the X chromosome from the father's sperm arrives in a 'pre-inactivated' state, which is a continuation of MSCI, and which persists during the period before the early embryo implants in the uterus. Okamoto *et al.*², however, now show that genes on the paternal X chromosome are transcriptionally

active at the very earliest embryonic stages, and that subsequent inactivation of the paternal X can occur without prior MSCI. This work shows that the two processes (MSCI and *de novo* paternal X inactivation after fertilization) can be mechanistically separated, and it confirms that there is a period after fertilization when the paternal X chromosome is transcriptionally active.

Both imprinted and random X-chromosome inactivation depend on the expression of a non-coding RNA called *Xist*. As the cells differentiate, *Xist* synthesis increases on the chromosome that is to be inactivated, coating the X chromosome and leading to the acquisition of repressive epigenetic marks (including modifications to core histones and to DNA) and gene silencing⁸. Okamoto and colleagues² took advantage of transgenic mice in which large pieces of the X chromosome, encompassing the *Xist* gene, had been inserted into non-sex chromosomes. They then studied various characteristics associated with X-chromosome activity and repression on both maternally and paternally inherited transgenic X sequences, and on the normal X chromosomes themselves. In all cases, the expression and dynamic epigenetic states of the transgenes were the same as those on the respective X chromosomes.

Okamoto *et al.* found that RNA signals —

representing gene transcription — occurred on both the maternal and paternal X-chromosome sequences in embryos at the two-cell stage (the stage at which the embryonic genome begins to be transcribed). Because the authors could not look at a large number of genes, it may still be that some genes on the paternal X are repressed at the two-cell stage — but these experiments clearly show that many are not.

Among the genes expressed on the paternal X at the two-cell stage is *Xist*. Shortly afterwards, sequences on the X chromosome that is to be inactivated undergo repressive modifications in a hierarchical and dynamic manner. For example, at the 8–16-cell stage, X-linked genes show a gradient of inactivation along the chromosome, with those farther from *Xist* being less tightly silenced⁶. These observations support a model in which the expression of *Xist* triggers an epigenetic inactivation process that spreads along the X chromosome during pre-implantation development. By contrast, the maternal copy of *Xist* is kept silent in the early embryo by an unknown epigenetic mark that originates in the egg⁹, thus keeping the maternal X active.

But there is one important difference between the transgenes and the normal X chromosome. The paternally inherited X has been subject to MSCI whereas the transgenic sequences have not. The mechanism of MSCI during sperm formation does not involve *Xist*, but a different epigenetic pathway⁵. It is unclear to what extent X-linked genes inactivated by MSCI remain silent during the later stages of sperm formation¹⁰, although it has been suggested that this silencing is carried over into the fertilized egg and gives rise to imprinted X inactivation^{6,7}. But the fact that the paternally inherited X transgenes are

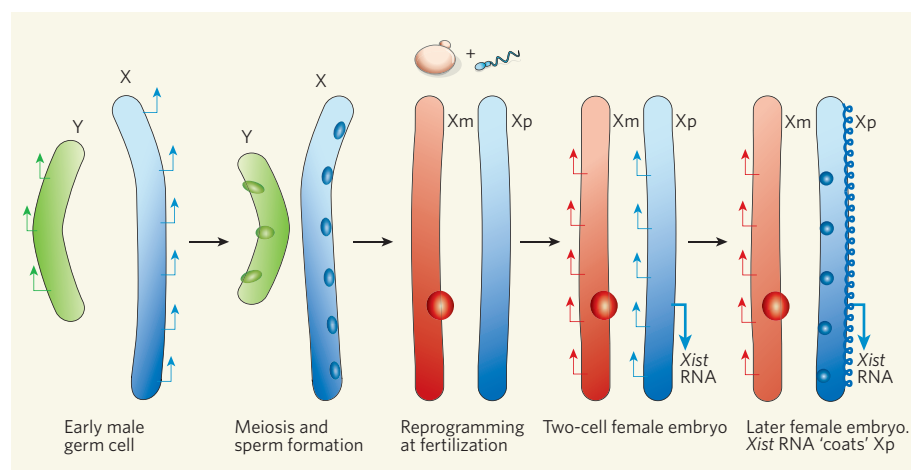


Figure 1 | Imprinted X inactivation: on again, off again? In early male germ cells, genes on the X and the Y chromosomes are transcribed (coloured arrows). During male meiosis, genes on the X and the Y become silenced and epigenetically marked (ovals). A female fertilized egg inherits the paternal Xp from the sperm and a maternal Xm from the egg. During or after fertilization of an egg, the epigenetic marks on the paternal X are reprogrammed; the *Xist* gene on the maternal X has been tagged with a repressive epigenetic mark before fertilization (circle). In the two-cell female embryo, genes on both Xs are transcribed, with the exception of the maternal *Xist* gene, which is repressed by the epigenetic mark. After the two-cell stage, *Xist* RNA begins to coat the paternal X, leading to silencing and epigenetic modification of the paternal genes. Genes on the maternal X remain active.



50 YEARS AGO

An Outline of the Cancer Problem.

— It is almost as difficult to review a popular book on cancer as to write one, and for similar reasons. The subject ranges over such a vast canvas that one is always acutely aware of numerous important omissions and of the minor distortions which inevitably appear... One is compelled to stress those aspects most dominant in one's own experience, and Dr. I. Hieger naturally stresses the important role of the chemical carcinogens in etiology. He was in the team which first isolated from coal tar a pure chemical carcinogen, and he gives us the most interesting story of this pioneer discovery. Moreover, we have seen lately the firm linkage of exposure to tobacco and industrial smoke to the tremendously rising incidence of lung cancer... If it is shown that lung cancer is in fact due to exposure to chemical carcinogens in smoke, their practical significance in human cancer will become of greatly enhanced importance.

From *Nature* 19 November 1955.

100 YEARS AGO

American palaeontologists are becoming more and more strongly convinced of the decisive character of the evidence afforded by extinct faunas of a comparatively recent connection between South America, South Africa, and Australia. A short time ago, Dr. W. B. Scott... announced his opinion that the fossil Santa Cruz insectivore *Necrolestes* is closely allied to the South African *Chrysochloris*, and that this relationship indicated a connection between South Africa and South America. Now Mr. W. J. Sinclair... states unequivocally that *Prothylacinus* and the other marsupial-like carnivores of the Santa Cruz beds are true marsupials closely related to the Australian *Thylacine*... Mr. Sinclair considers himself justified in stating that... "a land connection between Patagonia and the Australian region existed not later than the close of the Cretaceous or the beginning of the Tertiary".

From *Nature* 16 November 1905.

inactivated after fertilization in the absence of prior MSCI clearly shows that key aspects of imprinted X inactivation can take place without it.

One interesting outcome of these studies^{2,6,7} is that the pre-inactivation and the *de novo* inactivation hypotheses can both be correct, as summarized in Fig. 1. X-linked genes in the sperm could arrive in the egg in an epigenetically inactive form. The egg cytoplasm could then exert an immediate response to reverse this state — perhaps as part of the genome-wide reprogramming events that affect the paternal genome in the newly fertilized egg¹¹. It is likely, given the results of Okamoto *et al.*, that most genes on the paternal X, including the *Xist* gene, are activated at the two-cell stage. But it seems that over the next few cell divisions they become inactivated again through the action of *Xist*, without necessarily requiring prior MSCI.

The latest work illustrates the amazing plasticity and dynamic nature of epigenetic programming and reprogramming in germ

cells and early embryos. Insight into these events will give us clues about how the genome functions in normal development and in disease. It may also eventually provide tools for treating diseases of genome malfunction. ■

Wolf Reik is in the Laboratory of Developmental Genetics and Imprinting, The Babraham Institute, Cambridge CB2 4AT, UK.

e-mail: wolf.reik@bbsrc.ac.uk

Anne C. Ferguson-Smith is in the Department of Anatomy, University of Cambridge, Downing Street, Cambridge CB2 3DY, UK.

e-mail: afsmith@mole.bio.cam.ac.uk

1. Lyon, M. *Nature* **190**, 372–373 (1961).
2. Okamoto, I. *et al.* *Nature* **438**, 369–373 (2005).
3. Takagi, N. & Sasaki, M. *Nature* **256**, 640–642 (1975).
4. Lucchesi, J., Kelly, W. & Panning, B. *Annu. Rev. Genet.* PMID:16120055 24 August (2005).
5. Turner, J. *et al.* *Curr. Biol.* **14**, 2135–2142 (2004).
6. Huynh, K. & Lee, J. *Nature* **426**, 857–862 (2003).
7. Huynh, K. & Lee, J. *Nature Rev. Genet.* **6**, 410–418 (2005).
8. Brockdorff, N. *Trends Genet.* **18**, 352–358 (2002).
9. Tada, T. *et al.* *Development* **127**, 3101–3105 (2000).
10. Wang, P. J., Page, D. C. & McCarrey, J. R. *Hum. Mol. Genet.* **14**, 2911–2918 (2005).
11. Reik, W. & Walter, J. *Nature Genet.* **27**, 255–256 (2001).

COMMUNICATIONS TECHNOLOGY

Chaos down the line

Rajarshi Roy

Chaos, goes conventional wisdom, can only be a malign influence in telecommunications. But a technique that uses chaotically varying signals to transmit information more privately may help it shed that bad-boy image.

Synchronization leads to communication — even when the signals used are chaotic. That is the lesson of a study by Argyris and colleagues on page 343 of this issue¹. It reports the successful transfer of digital information at gigabit rates by chaotically fluctuating laser light travelling through more than 100 kilometres of a commercial fibre-optic link around Athens, Greece (Fig. 1). The transmitter and receiver become harmonized in chaotic synchrony, allowing information to be reliably extracted at the other end — a result that brings us closer to exploiting the inherent advantages of chaos, rather than trying to eliminate it whenever it appears.

The phenomenon of synchronization in

periodic systems has been known since at least 1665, when Christiaan Huygens observed that pendulum clocks become synchronized when placed close to each other on a common support. Asian fireflies flashing together, flocks of geese flying in remarkable formations and pedestrians in lock-step on London's Millennium Bridge are illustrations of synchronization when large numbers of living creatures get together². But synchrony also arises in inert matter: lasers and masers both exploit the ability of large ensembles of atoms and molecules to harmonize their oscillations and emit light in coherence. The key to synchrony in such systems is that the individual elements



Figure 1 | Attic experiment. Argyris and colleagues¹ successfully used chaotic waveforms to transmit information over a distance of more than 100 kilometres in the telecommunications network of Athens (seen here from the Acropolis).

P. A. SOUDERS/CORBIS

are in some way coupled to each other, leading to the formation of — sometimes beautiful — patterns in space and time.

So can synchrony be observed in systems that behave chaotically? One of the hallmarks of chaotic systems that are isolated from one another is that their evolution diverges exponentially fast, even when their initial conditions are very similar. This is why predicting the weather more than a few days in advance is so hard. Therefore, the finding a couple of decades ago that suitably coupled chaotic systems could (under certain conditions) do exactly the opposite, and converge from distant initial conditions to synchronize their chaotic motions, came as a huge surprise^{3,4}.

The subsequent discovery that lasers can emit light in patterns that are chaotic in time and space clearly implied that remote lasers should be able to synchronize if they receive light from one another, whether through space or through an optical fibre. Experiments soon confirmed this, and with the exchange of only small amounts of light, too^{5,6}. The synchronization is sufficiently robust for information to be exchanged: if one of the chaotic laser systems, the transmitter, is perturbed by a message source, it can 'fold' that information into its own chaotic waveform, which it transmits to the input of the second laser system, the receiver. Meanwhile, the receiver's output is a synchronized replica of the transmitter's original, unperturbed chaotic waveform — so the receiver recovers the message as the difference between its input and output waveforms. (Analogously, a radio receiver that is tuned to a carrier frequency recovers information from perturbations of the amplitude or frequency of a periodic waveform.)

Argyris and colleagues¹ take a large stride towards showing that the method of transmitting and receiving information using chaotic waveforms can also work in the real world, without the stable conditions of the laboratory. Applying chaotic systems in real-world communications is a beguiling prospect, because a third party intercepting the signal would have difficulty extracting the information sent. Such security aspects of chaos-based communications admittedly need much further analysis. But as the authors point out, chaotic carrier waveforms offer privacy in a manner that could be complementary to and compatible with conventional software-based and quantum-cryptographic systems.

The remarkable features of the authors' work are the simplicity of their set-up — they use chaotic diode laser systems and instrumentation that are widely available off the peg — and their demonstration that information can be recovered with quite reasonable bit-error rates over a commercial fibre-optic link. The optical fibre used for the experiments in Athens was temporarily free of network traffic, but was still installed and connected to the switches of the network nodes. The authors measured the characteristics of the fibre, such

as its attenuation and chromatic dispersion, before the experiment. This allowed them, for example, to exactly counter the effects of dispersion by inserting an appropriate length of dispersion-compensating fibre at the beginning of the link. Three amplifiers were used, one at the transmitter, one 50 kilometres from the transmitter, and one at the receiver, followed by optical filters with bandwidths of around 1 nanometre — this respectively compensated for optical losses and removed spontaneous noise.

The scheme used by Argyris *et al.* exploited time-delayed feedback to generate high-dimensional, high-capacity chaotic waveforms at high bandwidths. This has turned out to be a most fruitful approach: the bit-rate limit of several gigabits in these experiments is set by the electronic and optical components used, and could become much higher with suitably designed systems. For instance, the authors' strategy is compatible with a technique known as wavelength multiplexing; this allows much higher bit rates to pass through a single fibre by transmitting light of many different wavelengths simultaneously.

The exciting possibilities revealed by these experiments¹ may be pursued in other directions that more fully exploit the possibilities available for communication using electromagnetic waves. The vector properties of light waves (their polarization) could be used to encode data⁷. Optical patterns that are chaotic

in time and space might also be used to communicate holographic information⁸ by generalized synchronization. Here, transmitter and receiver do not share identical synchronized dynamics; instead, the relationship between the two is given by a mathematical function, supplying an additional element of privacy.

The success of such developments will ultimately depend on our willingness to implement new ways of transporting optical signals, as well as on novel transmitters and receivers. The rewards could be considerable, not only in understanding the communication of information using chaotic physical systems. Such work could in the long term also help us to elucidate the workings of that most private of communication networks — the human brain. ■

Rajarshi Roy is in the Department of Physics and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA.

e-mail: rroy@glue.umd.edu

1. Argyris, A. *et al.* *Nature* **438**, 343–346 (2005).
2. Strogatz, S. *Sync: The Emerging Science of Spontaneous Order* (Hyperion, New York, 2003).
3. Afraimovich, V. S., Verichev, N. N. & Rabinovich, M. I. *Izv. Vyssh. Uchebn. Zaved. Radiofiz.* **29**, 1050–1060 (1986).
4. Pecora, L. M. & Carroll, T. L. *Phys. Rev. Lett.* **64**, 821–824 (1990).
5. Roy, R. & Thornburg, K. S. Jr *Phys. Rev. Lett.* **72**, 2009–2012 (1994).
6. Sugawara, T. *et al.* *Phys. Rev. Lett.* **72**, 3502–3505 (1994).
7. Van Wiggeren, G. D. & Roy, R. *Phys. Rev. Lett.* **88**, 097903 (2002).
8. Rogers, E. *et al.* *Phys. Rev. Lett.* **93**, 244102 (2004).

CELL BIOLOGY

Two pores better than one?

Arnold J. M. Driessen

The movement of proteins through a cell's membrane requires a dedicated molecular machine. A glimpse of this apparatus in action shows that it has two channels, and hints at how these pores might be regulated.

A cell's membrane bristles with proteins that sense and communicate with its environment, and the cell secretes other proteins to send messages farther afield. To reach their destination, these proteins must travel from the aqueous environment of the cytoplasm where they are synthesized, through the seemingly impenetrable boundary of the lipid membrane. To ease the proteins' emigration, cells use a specialized protein complex called a translocase to direct proteins across or into the membrane. On page 318 of this issue, Mitra *et al.*¹ report the structure of this remarkable complex caught in the act of inserting a newly synthesized protein into the membrane.

Proteins that must be secreted or inserted in the cell membrane are synthesized (translated) by membrane-bound organelles called ribosomes; they have a short sequence at one end (the amino terminus) that acts as an address

label to signal their final destination. The translocase consists of a protein-conducting channel (PCC) in the membrane that binds to the ribosome and passes newly synthesized proteins bearing the appropriate address label across or into the cell membrane. The translocase must provide an aqueous path across the membrane for hydrophilic protein segments, as well as a side opening to the membrane lipid phase to release hydrophobic protein segments into the membrane. The aqueous path must be tightly controlled, or valuable ions and molecules will leak out of the cell.

Mitra *et al.*¹ have used cryo-electron microscopy to determine the structure of the Sec translocase, from the bacterium *Escherichia coli*², that is associated with a ribosome in the process of synthesizing a protein called FtsQ. They reveal a unique strategy by which a translating ribosome assembles a PCC

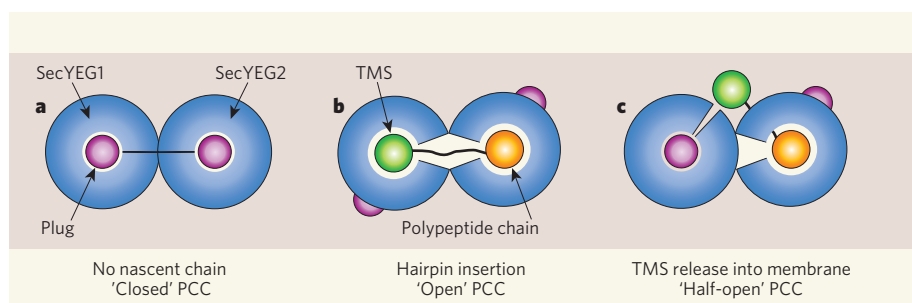


Figure 1 | Initial stages of the translocase mechanism, a model proposed by Mitra *et al.*¹. The protein-conducting channel (PCC) is shown as though looking through the pore from the inside of the cell. The membrane lipid is shown in light brown. The two SecYEG complexes of the PCC are shown as two clam-shell structures. The ribosome is not shown. **a**, In the 'closed' PCC structure, the channels are sealed by plug domains (purple). **b**, Insertion of a hairpin loop of nascent FtsQ protein with a transmembrane segment (TMS; green) and a hydrophilic segment (orange) will wedge open the SecY clam shells and displace the plugs to open a consolidated pore ('open' PCC). **c**, The ribosome rearranges the PCC, whereupon the inserted TMS is released sideways into the bulk lipid phase of the membrane from the SecYEG1 pore, which then closes. The SecYEG2 pore will remain open, but the channel will be sealed from the lipid phase by a wall formed by SecYEG1. The schematics of the closed and half-open PCC are based on the cryo-electron-microscopic structures described by Mitra *et al.*¹. The open PCC is hypothetical.

consisting of two separate pores that have different lipid accessibilities.

The Sec PCC is a complex of three proteins — SecY, SecE and SecG — each made of several transmembrane segments². Fully translated, partially folded polypeptide chains of the synthesized protein are pushed or pulled through the translocase channel with the aid of motor proteins — SecA in the case of the Sec translocase. But the translocase can also pass one end of a partially translated 'nascent' protein through, with the ribosome still at work at the other end churning out the growing peptide chain.

A previous structure³ of a closely related bacterial translocase from *Methanococcus jannaschii* shows that, when it is inactive, SecY resembles a clam shell, with each half of the clam consisting of five transmembrane segments. The SecY halves encompass a funnel-like cavity across the membrane that is blocked by a loop that forms a 'plug' at the external face. Biochemical evidence suggests that polypeptides travel through the centre of SecY⁴, and that during this time the plug is displaced⁵. On the inside face of this translocase, large loops extend out of the membrane plane, creating a docking site for the ribosome⁶.

On the basis of this structure, it was proposed that the active PCC consists of just one SecYEG complex, and that binding of the protein's address signal would displace the plug and force open the SecY halves to make a channel. However, several observations remained unexplained. For instance, if a single copy of the complex forms the PCC, why have several copies clumped together^{7,8} been observed biochemically and structurally? So Mitra *et al.*¹ reasoned that a snapshot of the PCC structure associated with a translating ribosome might provide a better insight. After some nifty experimental trickery to create a stable complex, they managed to catch the translocase red-handed, picturing the

ribosome, the peptide chain it was making and the PCC as one complex in unprecedented detail.

Strikingly, Mitra *et al.*¹ found that the ribosome-peptide-PCC structure contains two copies of the SecYEG complex. Computer modelling⁹ implied that the two SecYEG complexes face each other, with their lateral gates — the clam-shell mouths — touching (Fig. 1). This is in contrast to a recently proposed back-to-back arrangement^{2,8}, deduced from a structure obtained in the absence of any other proteins. However, several lines of evidence^{5,10} indicate that the back-to-back orientation of the PCC is probably an inactive, and possibly a resting, state of the translocase. The ribosome, and perhaps the SecA motor, might then help to ready the PCC for action by rearranging the individual SecYEGs into the face-to-face orientation.

Mitra and colleagues' structure shows three connections between the internal face of the PCC and the large subunit of the ribosome, including two that contact the large loops of the two SecY proteins — so the ribosome is in the ideal position to direct the rearrangement of the PCC. The structure exhibits a large frontal opening (about 20 × 40 Å) through which the nascent peptide chain and the connections are accessible from the cytoplasm. Remarkably, each SecYEG complex retains its own separate channel, rather than joining up to make one big one. Moreover, the two channels have different architectures: one is accessible to lipids, so it would seem to be adapted for integrating nascent membrane proteins into the membrane; the other is inaccessible to lipids, making it suitable for the transport of hydrophilic regions of the nascent polypeptide chain.

This latter 'aqueous' pore is occupied by a rod-like shape that seems to correspond to the hydrophilic segment of FtsQ. The structural interpretation suggests that the hydrophobic transmembrane segment of FtsQ is in

the lipid phase near the side gate of the other SecY. Mitra *et al.*¹ propose that this might be the site where SecY interacts with other proteins (such as YidC) that could chaperone the transmembrane segments into the membrane. The atomic model also predicts the position of the plug domain that closes the individual pores. In the 'aqueous' SecY channel, the plug appears in its open-state position at the periphery of SecY; but in the other channel, the plug seems to block the exit to the outside of the cell.

Nascent proteins destined for the membrane insert into the PCC as a hairpin, but how this occurs is still speculation. In the cryo-electron microscopy structure, the hairpin of the nascent chain has already slotted into the PCC, so it can provide few answers. Clearly, at some stage a single consolidated pore needs to form to direct the hydrophobic signal sequence or transmembrane segment through the lateral gate into the membrane (Fig. 1b), requiring opening of the 'mouths' of both SecYEG complexes. Once the hairpin has been inserted, the structure might then close up to generate the separate channels. The consolidation and division of the pore are probably controlled by the ribosome, but the mechanistic details remain to be elucidated. Likewise, the PCC structure suggests that the diameter of the consolidated pore may be varied to modify how far the SecYEG mouths open. If a transmembrane segment of the nascent chain is indeed reoriented in the PCC, a consolidated pore would provide sufficient manoeuvring space. Such a pore would also be wide enough to allow partially folded polypeptides through.

Mitra *et al.* present the first snapshot of a translocase in action, taking a significant step towards understanding the highly dynamic organization of this remarkable machine and how it controls distinct functions such as protein translocation and membrane-protein insertion. We eagerly await further developments in elucidating how the PCC is controlled by the motor protein SecA¹¹, which replaces the ribosome to push fully formed peptide chains through the membrane. ■

Arnold J. M. Driessen is in the Department of Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, and the Materials Sciences Center Plus, University of Groningen, 9751 NN Haren, The Netherlands. e-mail: a.j.m.driessen@rug.nl

1. Mitra, K. *et al.* *Nature* **438**, 318–324 (2005).
2. Osborne, A. R., Rapoport, T. A. & van den Berg, B. *Annu. Rev. Cell Dev. Biol.* **21**, 529–550 (2005).
3. van den Berg, B. *et al.* *Nature* **427**, 36–44 (2004).
4. Canon, K. *et al.* *J. Cell Biol.* **169**, 219–225 (2005).
5. Tam, P. C., Maillard, A. P., Chan, K. K. & Duong, F. *EMBO J.* **24**, 3380–3388 (2005).
6. Cheng, Z., Jiang, Y., Mandon, E. C. & Gilmore, R. J. *Cell Biol.* **168**, 67–77 (2005).
7. Manting, E. H. *et al.* *EMBO J.* **19**, 852–861 (2000).
8. Breyton, C. *et al.* *Nature* **418**, 662–665 (2002).
9. Tama, F., Miyashita, O. & Brooks, C. L. *J. Struct. Biol.* **147**, 315–326 (2004).
10. Veenendaal, A., van der Does, C. & Driessen, A. J. M. *J. Biol. Chem.* **276**, 32559–32566 (2001).
11. de Keyser, J. *et al.* *J. Biol. Chem.* **280**, 35255–35360 (2005).

BRIEF COMMUNICATIONS

Regional commitment to reducing emissions

Local policy in the United States goes some way towards countering anthropogenic climate change.

The non-participation of the United States in the recently ratified Kyoto Protocol¹ is a matter for global concern because it is estimated that the country produces 24% of all greenhouse-gas emissions worldwide². Here we analyse the commitment of individual states and municipalities to addressing this problem and find that, despite the federal policy, between 24 and 35% of the US population are currently (or soon will be) engaged in policies directed towards significantly reducing anthropogenic climate change. The importance of this sub-national effort, which we estimate corresponds to 27–49% of the gross domestic product, will depend — like the targets adopted in Kyoto — on the real reductions achieved in greenhouse-gas emissions.

The current administration plans to adopt carbon-intensity targets that would allow the United States a 30% increase in emissions above those specified in Kyoto's designated base year of 1990 (ref. 3). We have analysed the extent of the commitment at the sub-national level towards targets more like those of Kyoto. Because of the varying nature and maturity

of any such regional policies, we divided the states and municipalities adopting them into three categories — current, probable and possible 'adopters'. Current and probable adopters have climate-change policies that are similar in scope to Kyoto's recommendations; possible adopters may have targets commensurate with the Kyoto recommendations, but to be included in this category only pledges to reduce emissions are needed. Our analysis, which includes contributions by different states, counties and cities (for details, see supplementary information), indicates that the current adopters represent about 24% of the US population and contribute about 27% of the gross domestic product (GDP) (Table 1).

There are several limitations to climate-change policies that operate at the sub-national level, including a lack of mechanisms for enforcing such small-scale initiatives. The Kyoto agreement is now legally binding, but it too could be undermined if targets are continually postponed and if the threat of exclusion from a trading system that is not yet proven⁴ turns out to be ineffective. Nonetheless, many Kyoto signatories are taking real steps towards compliance. In the United States at the sub-national level, compliance will be a challenge even for current adopters, who have increased their carbon dioxide emissions on average by 14% since 1990 (ref. 5).

There is also the problem of carbon leakage (the process of moving high-emission activities across political borders), which affects Kyoto-ratifying countries as well. Carbon leakage from

countries required to cut emissions to those countries without targets has been estimated to be 6–25% (ref. 6), although more recently it has been suggested that carbon leakage under Kyoto could be greater than 100% (ref. 7). Comparable leakage could occur as a result of California's plan to become better connected to electrical grids with coal-burning states such as Utah and Nevada.

Although sub-national climate-change policies may suffer from limitations similar to those of the Kyoto Protocol, they could be more adaptable to regional heterogeneity in creating site-specific plans for reducing greenhouse-gas emissions. For example, US states have strong control over land use and agricultural policy, and moderate control over electrical systems, all of which are effective tools for mitigation of greenhouse gases⁸. Individual states could also help in reducing transportation emissions, although scant attention has been paid to this so far. In many states (including all the current adopters), transportation is the greatest contributor to greenhouse-gas emissions, and this issue will need to be addressed before directives can become properly effective.

The sub-national movement analysed here is an important deviation from US federal policy in that it acknowledges the need for climate-change policies and for setting targets. Pledges may be broken and adopting policies does not necessarily change behaviour. But it is encouraging that targets set over the past few decades to achieve environmental goals have been effective in driving change⁹ and have already helped to identify easy and inexpensive routes to emission reductions¹⁰.

Given that the US delegation to the next international climate conference may be excluded from many sessions¹¹, it is important that sub-national efforts persist. Even our lower-bound estimate of the amount of US



Downtown Los Angeles: California is one state driving emission cuts.

AP PHOTO/C. PIZZELLO

Table 1 | Sub-national contributors to greenhouse-gas reduction

	Population (millions)	Percentage of US population	Gross product in 2003 (US\$billions)	Percentage of US GDP
Current adopters				
California*	35.484	12.19	1,446	13.26
Connecticut†	3.483	1.20	172	1.58
Maine†	1.306	0.45	41	0.38
Massachusetts†	6.433	2.21	297	2.73
New Hampshire†	1.288	0.44	49	0.45
New Mexico‡	1.875	0.64	57	0.52
New York†	19.190	6.59	822	7.53
Rhode Island†	1.076	0.37	40	0.36
Vermont†	0.619	0.21	21	0.19
Subtotal	70.755	24.31	2,945	26.99
Probable adopters				
New Jersey	8.638	2.97	397	3.64
Oregon	3.560	1.22	120	1.10
Washington	6.131	2.11	245	2.24
Subtotal	18.329	6.30	763	6.99
Possible adopters				
25 US municipalities	12.774	4.38	1,673	15.34
Total	101.859	34.99	5,381	49.32

Percentage population figures are calculated for individual states with respect to the total US population of 291 million; the contribution of each state to the national economy is also shown, and is calculated from the 2003 value of the gross domestic product (GDP) of the United States (\$10.9 trillion). For details, see supplementary information.

*Pledged to reach 1990 levels by 2020. †Pledged 10% reduction below 1990 levels by 2020. ‡Pledged 10% reduction below 2000 levels by 2020.

GDP being pledged towards significant reductions (27%; Table 1) represents 9% of the global economy; the upper bound represents 16.7%, which is slightly larger than the GDP of Japan, the world's second largest economy.

Although there is no US federal cooperation with Kyoto, the implementation of climate-change policies by lower levels of government are widespread and governed by pledges that are not dissimilar from the targets adopted in Kyoto. These pooled efforts will ultimately be gauged by the real reductions in emissions that they achieve.

Brendan Fisher, Robert Costanza

Gund Institute for Ecological Economics,
Rubenstein School of Environment and Natural
Resources, University of Vermont, Burlington,
Vermont 05405, USA
e-mail: brendan.fisher@uvm.edu

- Butler, D. & Schiermeier, Q. *Nature* **436**, 156–157 (2005).
- Marland, G., Boden, T. & Andres, R. in *Trends: A Compendium of Data on Global Change* (Oak Ridge Natl Lab., Tennessee, 2003) http://cdiac.esd.ornl.gov/trends/emis/tre_usa.htm.
- Pew Center on Global Climate Change *Analysis of President Bush's Climate Change Plan* www.pewclimate.org/policy_center/analyses/response_bushpolicy.cfm (2002).
- Nature* **431**, 613 (2004).
- Emissions of Greenhouse Gases in the United States 2003* (Dept Energy, Washington DC, 2003).
- Babiker, M., Maskus, M. & Rutherford, K. *Carbon Taxes and the Global Trading System* (working paper 97-7) (Univ. Colorado, Boulder, 1997).
- Babiker, M. H. J. *Int. Econ.* **65**, 421–445 (2005).
- Pew Center on Global Climate Change www.pewclimate.org/policy_center/policy_reports_and_analysis/state/index.cfm (2004).
- Subak, S. E. *Nature* **374**, 300 (1995).
- Victor, D. G. & Salt, J. E. *Nature* **373**, 280–282 (1995).
- Haag, A. *Nature* **432**, 936 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/438301a

ANT NAVIGATION

Priming of visual route memories

Ants travelling to and fro between their nest and a foraging area may follow stereotyped foodward and homeward routes that are guided by different visual and directional memory sequences^{1–6}. Honeybees are known to fly a feeder-to-hive or hive-to-feeder vector according to whether or not they have recently fed — their feeding state controls which compass direction they select⁷. We show here that the feeding state of the wood ant *Formica rufa* also determines the choice between an outward or inward journey, but by priming the selective retrieval of visual landmark memories.

We trained the ants along a foraging route in which the appearance of a landmark differed on the ants' foodward and homeward paths. The ants ran 1 m from a start-pot to a drop of sucrose, both of which lay 20 cm from a black wall that was 2 m long and 20 cm high. They

were guided by the remembered appearance of the wall⁸, which was to their left on the way to food and to their right on the way home. (For methods, see supplementary information.)

To investigate the role of feeding state in priming visual memories for the foodward or homeward route, trained ants that had either been fed or left unfed were placed individually in a start-pot midway along the wall (Fig. 1a). Unfed ants walked so that they viewed the wall on their left (59 out of 63 paths from 23 ants); fed ants viewed the wall on their right (56 of 61 paths from 20 ants). The wall was regularly rotated during training and the visual scene was identical for fed and unfed ants. We conclude that the ants' feeding state, rather than the compass orientation or panoramic context, determines whether foodward or homeward memories are primed.

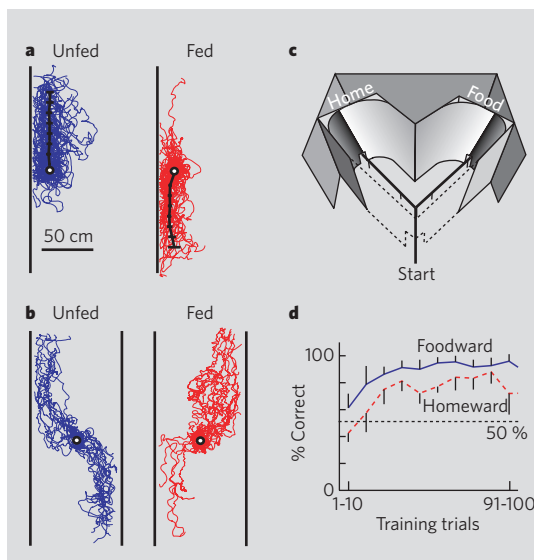


Figure 1 | Wood ants use feeding state to select visual memories for guiding routes towards food or the nest.

a, Individual trajectories of unfed and previously fed ants released from a start-pot (white circle). Thick line indicates the mean path, with 95% confidence interval (CI) plotted every 10 cm. **b**, Trajectories of ants when the start-pot is placed midway between two walls. The exit from the start-pot is at the top of the figure. Fed ants left in the direction of the exit; unfed ants circled the pot before choosing a direction. **c**, Y-shaped maze (dotted lines indicate front wall) with different patterns for homeward and foodward routes (see supplementary information). Distance from Y junction to pattern, 33 cm. **d**, Ants' performance on foodward and homeward journeys in training, showing mean choices of 11 ants (95% CI).

Ants did not just learn fixed motor patterns (for example, to turn left when fed) as the paths were the same, irrespective of whether the exit from the ants' start-pot faced the wall (as in training) or faced away. Ants placed midway between two identical walls 80 cm apart (Fig. 1b) walked closer to the left wall when unfed (20 of 20 paths, 9 ants) and to the right wall when fed (22 of 22 paths, 9 ants), thereby matching the route-specific visual memory that was primed by their feeding state.

In a second experiment, ants learned a foraging route in which they ran twice through the same 'Y' maze, first to reach food and then to be returned to the nest. Each arm of the 'Y' led to a different visual pattern (Fig. 1c). One of the two patterns signalled the way to food, and the other the way home. Training patterns were frequently switched between sides. Ants learned to choose the foodward pattern when unfed and the homeward pattern after feeding (Fig. 1d).

In unrewarded tests, unfed or previously fed ants made two journeys through the maze. On both journeys, ants chose the foodward pattern when unfed (first journey, 41 out of 44 correct; second, 37 of 41 correct) and the homeward pattern when fed (first journey, 38 of 43 correct; second, 25 of 28 correct) (all $P < 0.0001$). We conclude that the ants' visual memories were primed by feeding state and not by the sequence of rewarded patterns.

Although ants are rigid in sticking to familiar routes, they are like honeybees in that they show flexibility in choosing between routes. The selective priming of visual and vector memories specific to a particular route is an important component of this flexibility^{7,9–13}.

Robert A. Harris, Natalie Hempel de Ibarra, Paul Graham, Thomas S. Collett

School of Life Sciences, University of Sussex,
Falmer, Brighton, BN1 9QG, UK
e-mail: t.s.collett@sussex.ac.uk

- Santschi, F. *Rev. Suisse Zool.* **21**, 347–425 (1913).
- Collett, T. S., Dillmann, E., Giger, A. & Wehner, R. *J. Comp. Physiol. A* **170**, 435–442 (1992).
- Wehner, R., Michel, B. & Antonsen, P. *J. Exp. Biol.* **199**, 129–140 (1996).
- Collett, M., Collett, T. S., Bisch, S. & Wehner, R. *Nature* **394**, 269–272 (1998).
- Bisch-Knaden, S. & Wehner, R. *J. Comp. Physiol. A* **189**, 181–187 (2003).
- Kohler, M. & Wehner, R. *Neurobiol. Learn. Mem.* **83**, 1–12 (2005).
- Dyer, F. C., Gill, M. & Sharbowski, J. *Naturwissenschaften* **89**, 262–264 (2002).
- Graham, P. & Collett, T. S. *J. Exp. Biol.* **205**, 2499–2509 (2002).
- Wahl, O. Z. *Vergleich. Physiol.* **16**, 529–589 (1932).
- Koltermann, R. Z. *Vergleich. Physiol.* **75**, 49–68 (1971).
- Menzel, R., Geiger, K., Joerges, J., Müller, U. & Chittka, L. *Anim. Behav.* **55**, 139–152 (1998).
- Reinhard, J., Srinivasan, M. V., Guez, D. & Zhang, S. W. *J. Exp. Biol.* **207**, 4371–4381 (2004).
- Beugnon, G., Lachaud, J.-P. & Chagné, P. *J. Insect Behav.* **18**, 415–432 (2005).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/438302a

BRIEF COMMUNICATIONS ARISING online
♦ www.nature.com/bca see Nature contents.

ORIGIN OF FLIGHT

Could 'four-winged' dinosaurs fly?

Arising from: X. Xu *et al.* *Nature* **421**, 335–340 (2003); F. Zhang & Z. Zhou *Nature* **431**, 925 (2004)

Our understanding of the origin of birds, feathers and flight has been greatly advanced by new discoveries of feathered non-avian dinosaurs, but functional analyses have not kept pace with taxonomic descriptions. Zhang and Zhou describe feathers on the tibiotarsus of a new basal enantiornithine bird from the Early Cretaceous of China¹. They infer, as did Xu and colleagues from similar feathers on the small non-avian theropod *Microraptor* found in similar deposits², that these leg feathers had aerodynamic properties and so might have been used in some kind of flight.

Claims that these were 'four-winged' dinosaurs that took to the air require investigations using independent lines of evidence, which was done by neither team of authors. There is no evidence to confirm the existence of a four-winged stage that involved feathers on the hindlimbs and that could have been an intermediate between feathered, ground-dwelling theropods and aerial birds.

First, the long leg feathers in question are said to preserve either asymmetrical vanes² or little asymmetry but curved shafts¹, from which the authors infer that these feathers had a residual aerodynamic function¹. Neither feature indicates a function in lift or thrust (and therefore in flight). It has been suggested³ that both features are functionally correlated to flight, but the connection is not causal. Asymmetrical feathers resist torsion, and both features slightly reduce drag, which would be useful to a running animal with long feathers but would make no difference at realistic running speeds. Small non-avian theropods were good runners, but there is no evidence as yet that they flew.

Second, no evidence^{1,2} of how the feathers were attached to the hindlimb has been documented, so there is no indication that the feathers could have supported an aerodynamic load.

Third, Xu *et al.*² offer no support for their statement that the hindlimbs (and therefore the inferred 'hindwing') of *Microraptor* extended laterally like those of bats, but unlike those of any bird or other theropod dinosaur. Furthermore, no explanation is given of how the leg could have been raised and extended into the lateral plane without dislocating the hip joints. Their remarkable assertion that the tibia of *Microraptor* is 'bowed' is also unsupported, and this is more likely to be the result of being crushed and of crossing other preserved bones. Even a revised posture of the hindlimbs does not mitigate the case for 'four-winged gliding' — it would also have to be subjected to functional tests.

Fourth, the proposed function of leg feathers

in flight is unknown in birds of today, and the necessary structural modifications are not present in these extinct animals or in their relatives; therefore, the inferences drawn by the authors have no support from either the extinct or extant phylogenetic brackets⁴. Because leg feathers first appeared in non-flying theropods², some of considerable size, their original function was not in flight, and no flight-related explanation for their presence in basal birds is necessary⁵. Long leg feathers are found in owls and grouse, and in nearly all raptors⁶: as with the feathers in these extinct forms, these are long, vaned and nearly symmetrical, but they have no demonstrated role in flight.

Fifth, the analogy that the authors draw¹ between the use of webbed feet in kittiwakes as airbrakes and the extension of the hindlimbs in rapidly descending vultures¹ is inappropriate: the first example induces drag, and the second relocates the centre of mass as the bird increases the angle of attack of the wings. Neither function uses or requires leg feathers, or produces lift. To be useful in flight, the advantage to lift must outweigh the disadvantage of drag.

Sixth, the authors should have calculated the mass, wing area, centre of mass and the centre of lift that would have been necessary to test whether the reconstruction of *Microraptor* was likely to be able to glide². Such basic calculations must precede any assertion about gliding ability.

Most important, none of the evidence reported by the authors^{1,2} should have been

advanced as support for an 'arboreal' origin of bird flight. It is recognized that the arboreal versus cursorial dichotomy of models for the origin of bird flight is not capable of resolution and should have been abandoned long ago^{7,8}. Rather, the origin of the flight stroke is the central problem in the origin of flight, and so far nothing has been brought to light to indicate that *Microraptor* has any bearing on this question. Functional inferences should be based on functional studies, not just on analogies⁹. Ideas about arboreality should hinge on more than toe and claw proportions, and ideas about gliding on more than the profile of leg feathers.

Kevin Padian*, Kenneth P. Dial†

*Museum of Paleontology, University of California, Berkeley, California 94720, USA
e-mail: kpadian@berkeley.edu

†Division of Biological Sciences, University of Montana, Missoula, Montana 59812, USA

1. Zhang, F. & Zhou, Z. *Nature* **431**, 925 (2004).
2. Xu, X. *et al.* *Nature* **421**, 335–340 (2003).
3. Feduccia, A. & Tordoff, H. B. *Science* **203**, 1021–1022 (1979).
4. Witmer, L. D. in *Functional Morphology in Vertebrate Paleontology* (ed. Thomason, J. J.) 19–33 (Cambridge Univ. Press, 1995).
5. Gould, S. J. & Vrba, E. S. *Paleobiology* **8**, 4–15 (1982).
6. Perrins, C. M. *The Illustrated Encyclopedia of Birds* (Prentice Hall, New York, 1990).
7. Padian, K. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. A. & Gall, L. F.) 255–272 (Yale Univ. Press, New Haven, 2001).
8. Dial, K. P. *Science* **299**, 402–404 (2003).
9. Padian, K. *Am. Zool.* **41**, 598–607 (2001).

doi:10.1038/nature04354

ORIGIN OF FLIGHT

Xu *et al.* reply

Replying to: K. Padian & K. P. Dial *Nature* **438**, doi:10.1038/nature04354 (2005)

We agree that a strict biomechanical analysis is needed to reconstruct *Microraptor*'s locomotory mode, but we disagree with several of Padian and Dial's arguments¹. In addition to the six *Microraptor* specimens we described², other similarly preserved specimens³ have been discovered that also had long, asymmetrical pennaceous feathers attached to the hindlimbs². These feathers show features that are functionally correlated with flight⁴. A large, feathery surface on the legs would increase, rather than decrease¹, drag during running, as evidenced by the reduced or lost filamentous integumentary structures on the lower legs

of cursorial birds and mammals.

We have proposed a possible flight posture for *Microraptor*⁵, in which the hindlimbs stretch posteriorly, avoiding dislocation of the femur from the hip joint and maintaining consistency with the parasagittal posture of hindlimbs among dinosaurs: the hindlimb and tail feathers together form an aerofoil that provides a lift-surface, and the front wings supply the other lift-surface and generate thrust. We drew no functional inference from the bowed tibia^{2,5}.

Padian and Dial have overlooked a fundamental difference between the long leg feathers

of *Microraptor* and those of some living birds. The former are asymmetrical and have an arrangement similar to the flight feathers in the wings of flying birds², whereas the latter are bunched pennaceous feathers without aerodynamic modification. These functionally different structures are unsuitable for analysis by the extinct or extant phylogenetic bracket method.

The flight apparatus of *Microraptor*, such as the elongate, robust winged arms capable of flapping flight, is almost identical to that of *Archaeopteryx*. Some flight-related features, such as the large sternum and the relatively long flight feathers (compared with the forelimbs) are more developed in *Microraptor*. Assertions that *Microraptor* was non-flying on the basis of its non-avian status are unfounded without biomechanical analysis.

It is unwise to rely on the highly modified features of extant animals to deduce the behaviour of an ancient relative, without considering the basal taxon's morphology. For example, there was a stage in their evolution when whales had hindlimbs with large feet that functioned as flippers in the aquatic environment, before they became smaller and were

eventually lost⁶: without discovery of these basal forms, the extinct or extant phylogenetic bracket method would have predicted a different evolutionary pathway.

The systematic position of *Microraptor* as a close relative of birds and its possession of nearly all the flight features of basal birds make it ideal for studying the origin of flight, including the origin of the flight stroke. We disagree with Padian and Dial¹ that the arboreal versus cursorial debate is dead: either hypothesis could be accepted or rejected based on whether protoavians take advantage of gravity when taking off, which could be tested against a biomechanical analysis. *Microraptor*, which provides negative evidence for the cursorial hypothesis², could have taken to the air⁷ according to our recent biochemical analysis. Moreover, a new maniraptoran of the avian lineage has long metatarsalian feathers, indicating that hindwings may have been a common adaptation during the dinosaur–bird transition⁸. This unusual morphology should not be ignored in the study of the origin of flight, particularly as it bears distinct aerodynamic features and occurs phylogenetically

close to the origin of birds.

Xing Xu[†], Zhonghe Zhou*, Xiaolin Wang*, Xuewen Kuang‡, Fucheng Zhang*, Xiangke Du§

*Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, PO Box 643, Beijing 100044, China

†Present address: American Museum of Natural History, New York, New York 10024, USA

e-mail: xu@amnh.org or xing_xu@sina.com

‡Tianjin Museum of Natural History, Tianjin 300074, China

§Radiological Department, People's Hospital, Beijing University, Beijing 100044, China

1. Padian, K. & Dial, K. P. *Nature* **438**, doi:10.1038/nature04354 (2005).
2. Xu, X. et al. *Nature* **421**, 335–340 (2003).
3. Ji, Q. et al. in *The Mesozoic Jehol Biota in Western Liaoning, China* 1–373 (Geology Press, Beijing, 2004).
4. Feduccia, A. & Tordoff, H. B. *Science* **203**, 1021–1022 (1979).
5. Xu, X., Zhou, Z., Zhang, F., Wang, X. & Kuang, X. *J. Vert. Paleontol.* **24** (suppl.), 133A (2004).
6. Theewissen, J. G. M., Hussain, S. T. & Arif, M. *Science* **263**, 210–212 (1994).
7. Dyke, G. J., Nudds, R. M., Rayner, J. M., Norell, M. & Xu, X. *Proc. R. Soc. Lond. B* (submitted).
8. Xu, X. & Zhang, F.-C. *Naturwissenschaften* doi:10.1007/s00114-004-0604-y (2005).

doi:10.1038/nature04355

ORIGIN OF FLIGHT

Zhang & Zhou reply

Replying to: K. Padian & K. P. Dial *Nature* **438**, doi:10.1038/nature04354 (2005)

Padian and Dial¹ challenge our view that the evolution of flight involved a four-winged stage. This disagreement stems from our different views on the origin of bird flight and from the methodology we use to analyse functional morphology in the non-avian theropod *Microraptor*² and in an enantiornithine bird³ from the Early Cretaceous period in China.

Although the wing produced the main force for lift and thrust in the new enantiornithine bird, the curved shaft and asymmetric vanes of the leg feathers indicate that they could still produce some lift and thrust, as well as reducing drag. The arboreality of this bird is indicated by the curvature of the pedal claws and toe proportions⁴. It is not self-evident that small non-avian theropods were good runners: indeed, some of them may actually have been arboreal^{5,6}.

Contrary to Padian and Dial's assertions¹, attachments of leg feathers to the external side of the leg bones (tibia or metatarsals) have now been documented not only in *Microraptor gui*

but also in *Archaeopteryx*⁷, *Confuciusornis*⁸ and several enantiornithine birds^{3,8}, which would have enabled them to be extended into the lateral plane to produce some lift.

Padian and Dial argue that there is no evidence for a four-winged stage in flight origin. Leg feathers first appeared in non-flying theropods, which may have conferred some aerodynamic advantage, judging by their asymmetry and by the presence of an alula in the forelimb. So the original function of leg feathers might have been for gliding. We agree with Padian and Dial that these feathers were probably not used for active flight in *Microraptor* or in early birds. Although long leg feathers are found (but not to assist flight) in many living birds, these do not show the asymmetry or curved shaft indicative of aerodynamic properties.

The absence of aerodynamic function in leg feathers in modern birds does not mean that it did not exist in their ancestors. Reduction of drag is probably one of the main functions of

the leg feathers in *Microraptor* and early birds, and modern birds such as kittiwakes use the leg feathers on their extended hindlimbs as well as their webbed feet for this purpose.

Although we believe that leg feathers had some aerodynamic function in *Microraptor* and that this was weaker in early birds, they were probably not the main flight apparatus. However, their role in the origin of bird flight should not be ignored — functional inferences should be based mainly on functional studies, not on phylogenetic brackets⁹.

Zhonghe Zhou, Fucheng Zhang

Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China

e-mail: zhonghe@yeah.net

1. Padian, K. & Dial, K. P. *Nature* **438**, doi:10.1038/nature04354 (2005).
2. Xu, X. et al. *Nature* **421**, 335–340 (2003).
3. Zhang, F. & Zhou, Z. *Nature* **431**, 925 (2004).
4. Zhou, Z. & Farlow, J. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 237–254 (Yale Univ., New Haven, 2001).
5. Xu, X., Zhou, Z. & Wang, X. *Nature* **408**, 705–708 (2000).
6. Zhang, F., Zhou, Z., Xu, X. & Wang, X. *Naturwissenschaften* **89**, 394–398 (2002).
7. Christiansen, P. & Bonde, N. C. R. *Palevol.* **3**, 99–118 (2004).
8. Zhou, Z. & Zhang, F. *Vertebrata Palasiatica* (in the press).
9. Zhou, Z. *Naturwissenschaften* **91**, 455–471 (2004).

doi:10.1038/nature04356

REVIEWS

Potential impacts of a warming climate on water availability in snow-dominated regions

T. P. Barnett¹, J. C. Adam² & D. P. Lettenmaier³

All currently available climate models predict a near-surface warming trend under the influence of rising levels of greenhouse gases in the atmosphere. In addition to the direct effects on climate—for example, on the frequency of heatwaves—this increase in surface temperatures has important consequences for the hydrological cycle, particularly in regions where water supply is currently dominated by melting snow or ice. In a warmer world, less winter precipitation falls as snow and the melting of winter snow occurs earlier in spring. Even without any changes in precipitation intensity, both of these effects lead to a shift in peak river runoff to winter and early spring, away from summer and autumn when demand is highest. Where storage capacities are not sufficient, much of the winter runoff will immediately be lost to the oceans. With more than one-sixth of the Earth's population relying on glaciers and seasonal snow packs for their water supply, the consequences of these hydrological changes for future water availability—predicted with high confidence and already diagnosed in some regions—are likely to be severe.

Water is essential to human sustenance. Well over half of the world's potable water supply is extracted from rivers, either directly or from reservoirs. The discharge of these rivers is sensitive to long-term changes in both precipitation and temperature, particularly in the snowmelt-dominated parts of the world. Changes in the amount of precipitation tend to affect the volume of runoff and particularly the maximum snow accumulation, which usually occurs near the end of the winter at the onset of the melt season. On the other hand, temperature changes mostly affect the timing of runoff. Increasing temperatures lead to earlier runoff in the spring or winter, and reduced flows in summer and autumn—at least in the absence of changes in precipitation.

In general, the direction and (to a lesser extent) the magnitude of surface temperature changes are much more consistent among climate models than are precipitation changes¹. Near-surface air-temperature predictions from existing global climate models that are forced with anthropogenic increases in atmospheric greenhouse gas concentrations imply a high degree of confidence that future changes to the seasonality in water supply will occur in snowmelt-dominated regions. Even for models with temperature sensitivities near the lower end of the predicted range, impacts on snowmelt-dominated regional water resources are substantial². Indeed, such changes are already obvious in the observational records of key components of the hydrological cycle, such as snow pack in the western USA^{3–5}. Taken together, the predictions and observations portend important issues for the water resources of a substantial fraction of the world's population.

It is generally thought that increasing greenhouse gases will cause the global hydrological cycle to intensify¹, with benefits for water availability^{1,6}, although a possible exacerbation of hydrological extremes may counteract the benefits to some degree. However, in regions where the land surface hydrology is dominated by winter

snow accumulation and spring melt, the performance of water management systems such as reservoirs, designed on the basis of the timing of runoff, is much more strongly related to temperature than to precipitation changes. Even though there is relatively little agreement among the global models as to the magnitude (and even direction of) precipitation changes regionally^{7–10}, there is no indication for a seasonal shift of precipitation to the summer and autumn. The projected changes in temperature therefore strongly imply future changes of seasonal runoff patterns in snowmelt-dominated regions.

The hydrological cycle at the land surface includes the processes of snow/ice accumulation and melting as well as the impact these processes will have on regional changes in evaporative demand. In a warmer climate, snow will melt earlier in the year than it did before and in some places this has already happened^{3,11,12}. Taken together, these impacts mean less snow accumulation in the winter and an earlier peak runoff in the spring.

On a global scale, the largest changes in the hydrological cycle due to warming are predicted for the snow-dominated basins of mid- to higher latitudes, because adding or removing snow cover fundamentally changes the snow pack's ability to act as a reservoir for water storage¹³. Studies in various regions of the globe indicate that the stream-flow regime in snowmelt-dominated river basins is most sensitive to wintertime increases in temperature^{12,13}. Because of this, and also because there is little certainty in precipitation predictions^{7–10}, we focus here on the sensitivity of water resources in snowmelt-dominated regimes to temperature.

All models show warming with increasing greenhouse gases, so we can begin to say with some certainty how some critical components of the hydrological cycle will respond in the future.

Global distribution of snowmelt-dominated runoff

We used a spatially distributed macroscale hydrology model¹⁴ to identify the regions of the globe where snowmelt plays a dominant

¹Climate Research Division, Scripps Institution of Oceanography, La Jolla, California 92093, USA. ²Department of Civil and Environmental Engineering, ³Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington 98195-2700, USA.

role in the seasonal patterns of stream-flow. The model was run over all global land areas (excluding Antarctica and Greenland) at a spatial resolution of 0.5° latitude/longitude for a twenty-year (1980–1999) period. We approximated the importance of snow to annual runoff by using the ratio R of the accumulated annual snowfall to annual runoff (Fig. 1, colour scale). This allowed us to determine whether or not runoff for each grid cell is snowmelt-dominated by using the criterion that $R > 0.5$ for these cells.

We compared, for each of the world's major river basins, the simulated annual runoff to the estimated reservoir storage capacity^{15,16} in order to determine cases where reservoir storage capacity is adequate to buffer large seasonal stream-flow shifts (and hence exclude basins that, in spite of being snowmelt-dominated, would be insensitive to shifts in runoff timing). Watersheds within the snowmelt-dominated domain that meet these criteria include the Colorado River, the Churchill River and the Grand River (all in North America), and the Angara River (a tributary of the Yenisei River) in Asia. The red outline in Fig. 1 shows the domain where runoff is snowmelt-dominated minus the four basins identified as having large storage capacities relative to runoff. Within this domain, water resources are arguably susceptible to warming-induced shifts in stream-flow seasonality.

In general, the snowmelt-dominated regions occupy parts of the globe that are at latitudes greater than $\sim 45^\circ$ (North and South), with some exceptions. (1) Mountainous regions (except those nearest the Equator) are generally snowmelt-dominated (the inset of Fig. 1 shows the regions of the world that are topographically complex according to a criterion based on average slope¹⁷). (2) Some regions poleward of 45° North that are warmed by oceans do not experience enough snowfall to be snowmelt-dominated (for example, parts of Europe and the coastal regions of the USA Pacific Northwest and British Columbia). (3) Cold dry regions that experience little wintertime precipitation also do not receive enough snowfall to be snowmelt-dominated (for example, northeastern China).

The domain of influence within the red line of Fig. 1 is almost certainly underestimated, because the criterion we used is applied on a grid cell by grid cell basis, and does not account for areas where water availability is predominantly influenced by snowmelt that is generated upstream. Therefore, we extended the domain of influence

into sub-basins where the annual runoff originating in the snowmelt-dominated cells accounts for at least 50% of the runoff for the entire sub-basin (black lines in Fig. 1). These regions include parts of northern China, northwestern India, areas south of the Hindu Kush, sub-basins downstream of the southern Andes, north-central USA, and some coastal areas of western North America and Europe. According to a year 2000 population map¹⁸, approximately one-sixth of the world's population lives within this combined snowmelt-dominated, low-reservoir-storage domain. The population affected by warming-induced shifts in water availability is most probably greater than this estimate because we do not account for populations that derive their water resources from outside the basins in which they dwell. Note that the combined region in Fig. 1 encompasses much of the industrialized world, accounting for roughly one-quarter of the global gross domestic product.

Evapotranspiration in a warming climate

Our discussion so far has focused on the direct effects of warming on stream-flow seasonality in snowmelt-dominated regions. Warming-induced changes to evapotranspiration may also affect regional water availability. Unfortunately, there is little agreement on the direction and magnitude of historical, let alone one predicted, evapotranspiration trends. Observations from various countries in the Northern Hemisphere show that pan evaporation has been steadily decreasing for the past fifty years, contrary to the expectation that warming would cause increased evaporation^{19–22}. Two proposals exist to explain this paradox.

First, decreasing pan evaporation trends may be indicative of increasing actual (as opposed to potential) evapotranspiration in moisture-limited regions because increased land surface evaporation alters the humidity regime surrounding the pan, causing the air over the pan to be cooler and more humid^{23–26}. Second, consistent declines of pan evaporation, diurnal temperature range, and global solar irradiance suggest that actual evapotranspiration is also declining because of increased cloudiness and concentrations of atmospheric aerosols that systematically reduce surface energy availability for evaporation^{19,27–29}. Changes in wind speed or in the attenuation of wind at the surface due to changes in vegetation at observing sites

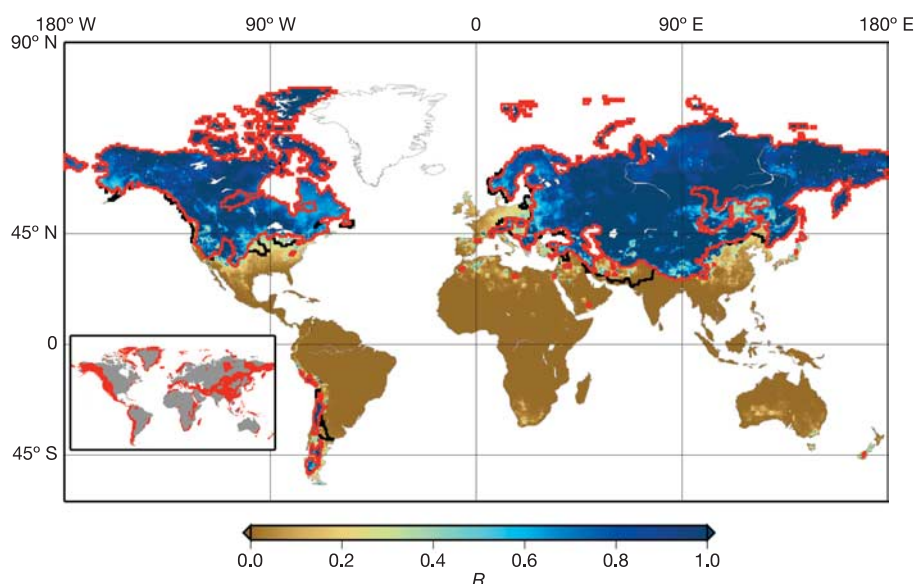


Figure 1 | Accumulated annual snowfall divided by annual runoff over the global land regions. The value of this dimensionless ratio lies between 0 and 1 and is given by the colour scale, R . The red lines indicate the regions where streamflow is snowmelt-dominated, and where there is not adequate reservoir storage capacity to buffer shifts in the seasonal hydrograph. The

black lines indicate additional areas where water availability is predominantly influenced by snowmelt generated upstream (but runoff generated within these areas is not snowmelt-dominated). The inset shows regions of the globe that have complex topography using the criterion of ref. 17.

may also play some role in apparent downward trends in pan evaporation data³⁰.

Ohmura and Wild²⁸ discuss some complications that impede our understanding of global trends in evapotranspiration. In snowmelt-dominated regions, though, these uncertainties are arguably of reduced importance, because changes in the timing of snowmelt runoff induce a negative feedback on changes in evapotranspiration. Earlier melt results in increased soil moisture (and so also the water available for evapotranspiration) earlier in the season, a time when potential evaporation (dominated by net radiation) is low. Later in the year, when potential evaporation is higher, the shift in snowmelt timing reduces soil moisture, and hence evaporative resistance is increased, again reducing the effect of evaporation changes. Therefore, although changes in evapotranspiration are critical to runoff production in most hydrological regimes, their effect (and hence the effects of the above-noted uncertainties) are attenuated in the snowmelt-dominated regions of the globe.

Impacts on regional water supplies

We examine three case studies from different parts of the world that are in the snowmelt-dominated domain. These case studies were selected to help provide an appreciation for the magnitude of the potential regional water problems that may be associated with shifts in the seasonality of runoff associated with climate change.

Western USA. The Accelerated Climate Prediction Initiative (ACPI)² demonstration project was launched in 2000 to investigate the impacts of greenhouse warming on water supplies in the western United States³¹. The methods and detailed results are included in 16 papers in a special volume of the journal *Climatic Change*². The most obvious signature of climate change in the simulations generated by this project was a general warming over the western USA: a warming that by the middle of the 21st century was projected to be 0.8–1.7 °C greater than present values. This warming is projected to be accompanied by little or no change in precipitation according to the climate change scenarios generated for the project by the NCAR-DOE Parallel Climate Model². In the western USA, much of the annual precipitation falls as snow in the mountains during the winter, and then melts during the spring and summer: that is, it is within the red lines shown in Fig. 1.

The most significant impact of a general warming was found to be a large reduction in mountain snow pack and a substantial shift in stream-flow seasonality, so that by 2050, the spring stream-flow maximum will come about one month earlier in the year. There is not enough reservoir storage capacity over most of the West to handle this shift in maximum runoff and so most of the 'early water' will be passed on to the oceans. These hydrological changes have considerable impacts on water availability and are discussed in the literature². For example, in the Columbia River system, less winter snowfall and earlier melting will force residents and industries to face, by 2050 or before, a choice of water releases for summer and autumn hydro-electric power or spring and summer releases for salmon runs. The ACPI research shows that, with the predicted climate change, the river cannot be managed to accommodate both, unless we are ready to accept substantial (10–20%) reductions of hydropower generation or serious harm to the federally protected salmon population of the region (Fig. 2)³².

The Rhine River in Europe. Climate-change simulations project a warming in the Rhine River basin of 1.0–2.4 °C over present values by the middle of the century¹. Hydrological simulations suggest that this warming will shift the Rhine River basin from a combined rainfall and snowmelt regime to a more rainfall-dominated regime, resulting in an increase in winter discharge, a decrease in summer discharge, increases in the frequency and height of peak flows, and longer and more frequent periods of low flow during the summer³³. Socio-economic implications include: a reduction in water availability for industry, agriculture and domestic use during the season of peak demand (which is further stressed by an increase in summertime

demand due to higher temperatures); an increase in the number of low-flow days during which ships cannot be fully loaded on major transport routes (causing an increase in transportation costs); a decrease in the level of flood protection (given no additional implementation of flood defence measures); a decrease in annual hydropower generation in some parts of the basin; and a loss in revenue due to a shortened ski season³³.

Canadian prairies. Climate studies for the Canadian prairies generally agree that a doubling of atmospheric CO₂ will result in an increase in surface air temperature (possibly as much as 8 °C during winter), a decrease in snow pack, an earlier snowmelt, and a decrease in summer soil moisture³⁴. These effects and a longer period of low flows during summer and autumn could lead to an increase in the frequency and severity of droughts³⁵. Historically, nearly 50% of the water use over the Canadian prairies has been for agriculture through irrigation, and this demand has been met primarily with surface water, unlike the prairies of the USA, which rely also on groundwater^{34,36}. For this reason and because stream-flows are limited and extremely variable from year to year, agriculture in the Canadian prairies is very sensitive to drought^{34,36}. Although global climate models do not predict great changes in precipitation for Canada, an earlier spring runoff peak will probably cause agriculture in the Canadian prairies to become more at risk in a warming climate³⁷. Furthermore, increased water demand for irrigation will also lead to heightened competition with other water needs, including stream-flow requirements to maintain aquatic habitat, and the needs of water users downstream of the Alberta–Saskatchewan border (under a 1969 agreement, Alberta must allow 50% of stream-flow to pass downstream of the border)³⁶.

Summary of regional impacts. The studies summarized above show that current demands for water in many parts of the world will not be met under plausible future climate conditions, much less the demands of a larger population and a larger economy.

The physics behind this statement is temperature-driven, not precipitation-driven, and this makes the conclusions robust because all current models predict a warmer future world. The other key factor affecting water availability is the lack of enough reservoir storage to manage a shift in the seasonal cycle of runoff. Current information about the climate-related water challenges facing much of the world, although by no means perfect, is sufficiently robust that major future problem areas can now be defined. The matter takes on a greater urgency because the model-predicted signals are already being observed.

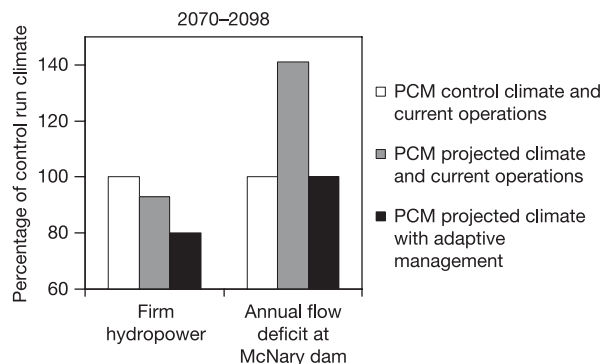


Figure 2 | Trade-off between firm hydropower and stream-flow requirements. The effect of Parallel Climate Model (PCM) climate change projections for the period of 2070 to 2098 on Columbia River Basin reservoir system reliabilities, as compared to the PCM control climate and operations scenario. Implementing adaptive management reduces the annual environmental flow deficit at McNary Dam in southeastern Washington, USA (benefiting salmon), but decreases firm (reliable) hydropower. Figure created by A. Hamlet using results from ref. 32.

Will changes in precipitation patterns offset the problems associated with warming? The most likely answer is 'no'. If less rain falls over a region, water availability will decrease. If more rain falls and the reservoir storage capacity is much less than the annual runoff, then the water will be lost downstream (to the ocean in many cases)—particularly in regions, like the western USA, where precipitation is mainly in winter and the effective storage capacity of winter snow pack will be lost. The changes in precipitation required to ameliorate the problem would have to come through a shift in the seasonal cycle of rainfall towards the dry season, a feature that is not usually exhibited by anthropogenically forced climate models.

Two examples of impacts on glaciers. The results for the regional water resources case studies discussed above and the simple physics behind them seem likely to be qualitatively reproduced in virtually all regions where snowmelt is important to local water availability⁶ and where annual runoff exceeds storage capabilities. Our results in the western USA suggest that even more serious problems may occur in regions that depend heavily on glacial meltwater for their main dry-season water supply. This is because, once the glaciers have melted in a warmer world, there will be no replacement for the water they now provide, in contrast to the present snow-pack-dependent water supply that is renewed seasonally. In this case, the natural storage of fossil water in the glaciers has even more importance than seasonal storage in just the snow pack. It is well documented that glaciers are in retreat over most (but not all) of the world^{1,38,39}, so the threat here seems both real and immediate—a situation also well documented in the world's press over the past several years.

Himalaya–Hindu Kush region. Perhaps the most critical region in which vanishing glaciers will negatively affect water supply in the next few decades will be China and parts of Asia, including India (together forming the Himalaya–Hindu Kush (HKH) region), because of the region's huge population (about 50–60% of the world's population). The ice mass over this mountainous region is the third-largest on earth, after the Arctic/Greenland and Antarctic regions. The hydrological cycle of the region is complicated by the Asian monsoon, but there is little doubt that melting glaciers provide a key source of water for the region in the summer months: as much as 70% of the summer flow in the Ganges and 50–60% of the flow in other major rivers^{40,41,42}. In China, 23% of the population lives in the western regions, where glacial melt provides the principal dry season water source⁴³.

There is little doubt that the glaciers of the HKH region are melting and that the melting is accompanied by a long-term increase of near-surface air temperature (ref. 44 and Figs 2.9 and 2.10 in ref. 1), the same level of warming we saw impacting the western USA. After 25 years of study, the China Glacier Inventory was recently released⁴⁵. It showed substantial melting of virtually all glaciers, with one of the most marked retreats in the last 13 years (750 m) of the glacier that acts as one of the major sources of the Yangtze River, the largest river in China. In total, it is estimated that the entire HKH ice mass has decreased in the last two decades. Furthermore, the rate of melting seems to be accelerating⁴⁶.

The few analytical studies that exist for the region suggest both a regression of the maximum spring stream-flow period in the annual cycle by about 30 days (ref. 47) and an increase in glacier melt runoff by 33–38% (ref. 48). These numbers seem consistent with what is being observed and bear striking similarities to the stream-flow results from the western USA. The huge inconsistency, however, occurs in the impacts on local water supplies. In the western USA, model-predicted impacts are already being seen in the hydrological cycle. The models suggest that the impacts will appear as a long-term trend in snow amount and runoff. But in the HKH region, there may (for the next several decades) appear to be normal, even increased, amounts of available melt water to satisfy dry season needs. The shortage, when it comes, will likely arrive much more abruptly in time; with water systems going from plenty to want in perhaps a few decades or less.

It appears that some areas of the most populated region on Earth are likely to 'run out of water' during the dry season if the current warming and glacial melting trends continue for several more decades. This may be enough time for long-term planning to see just how the region can cope with this problem. Unfortunately, the situation here is that when the glaciers melt and their fossil water is used or lost, their contribution to the water supply of the region will cease.

South American Andes. A large fraction of the population living west of the South American Andes relies on the glacial melt from those mountains to feed the area's rivers to supply water and hydro-power. Without the glacier-supplied river water, the people and economies of the region would have to undergo tremendous adjustments^{49,50}. The physics governing the Andean glaciers are more complicated than simple temperature forcing. Depending on the latitude and on which side of the Andes we consider, the glaciers' mass balance can be controlled by different factors^{51,52}. Although air temperature changes are still important in most areas, other processes (such as moisture flux and precipitation) dominate in some regions. This makes the prediction of what might happen in the Andes much more difficult. Although all greenhouse models predict warming air temperatures, they can disagree on predicted changes in rainfall, moisture flux, and so on.

In spite of this complexity, melting of the glaciers is well documented for the Andes^{53,54}. In Peru alone, the glacier-covered area has been reduced by 25% in the last three decades (as reported at the Conference on Mass Balance of Andes Glaciers, Huaraz, Peru, 6–9 July 2004; http://www.inrena.gob.pe/serusu/serusu_ppoint.htm). At current rates, some of the glaciers may disappear in a few decades, if not sooner. The high-frequency surges and retreats and the uneven spatial distribution of the general glacier retreat makes understanding and predicting the behaviour of glaciers in this area uncertain.

The melting started some decades ago. The International Panel for Climate Change (IPCC) shows a long-term trend in increasing air temperature in the region (ref. 38 and Figs 2.9 and 2.10 in ref. 1). Higher-resolution, more-detailed analysis of many stations in the region show a similar temperature increase, one that seems to be increasing^{55,56}. Consider the case of Quelccaya in the Andes (Fig. 3). When the summit core was originally drilled in 1976, it contained clear annual cycles in its layering that extended back in time for approximately 1,500 years (ref. 38). When it was re-drilled in 1991, the annual layers in the upper 20 m of the core had been obliterated by percolation of meltwater. Together, these two results show that melting at the summit had occurred, a condition that had not previously occurred in the last 1,500 years. The probability seems high that the current glacier melting in the Andes will continue, just as it will in Asia (and other regions of the world). It is fossil water that has been lost and will not be replaced anytime soon, especially not in the context of anthropogenically induced greenhouse warming. The results and projections suggest that current dry-season water resources will be heavily depleted once the glaciers have disappeared.

Some uncertainties in estimating impacts. All of the future climate predictions have uncertainties. We touch on only a few of the more important ones below, with the goal of seeing whether they might overcome the warming signal and make the conclusions above moot. We do not, however, attempt here a complete discussion of all the uncertainties that attend climate models.

In some cases, the uncertainties have to do with the models' inability to reproduce today's climate, casting doubt on future climate predictions. Predictions using regional, high-spatial-resolution models, of the type needed for regional water studies, are only now starting to come into their own in the greenhouse arena, but they carry a whole set of problems in addition to those associated with the coupled atmosphere–ocean general circulation models (CGCMs). For instance, they often have different physics from the CGCMs—there are scale-dependence issues, and new levels of parameterizations are required. However, such regional models will

be required for good quantitative estimates of potential future water problems. Such high-resolution, regional hydrological studies have not yet been undertaken for either the HKH region or South America.

One of the greatest uncertainties in future prediction has to do with how the models are forced. Stated more directly, what are the implications of omitting forcings that we strongly suspect (or know) are important but cannot yet reliably be included in the model physics? Of these, the most important is thought to be the incomplete inclusion of aerosols and their impacts, especially on clouds. Excellent discussions of the current state of the aerosol problem may be found in refs 57 and 58, and ref. 59 shows the sensitivity of climate model predictions to uncertainties in indirect aerosol forcing.

The key question for this paper is: Can the aerosol/cloud problem overwhelm the direct greenhouse-gas-induced temperature forcing that affects the regional hydrological cycle, giving net cooling as opposed to warming? We consider below some of these uncertainties qualitatively to see how they might impact the results discussed above.

Aerosols and clouds. Aerosols are thought to cool the planet's surface through increased scattering and cloud cover and re-radiation of solar energy to space. The representation of clouds in CGCMs carries a large uncertainty all by itself, but the joint interaction of clouds and

aerosols represents one of the major challenges to climate modellers today. Virtually all climate models have some representation of direct aerosol effects (that is, reflectivity of the particles) in them, but none have yet fully included the indirect effects (for example, the effect of aerosols on cloud distributions via their role as cloud condensation nuclei, or other effects discussed below). A preliminary study⁶⁰ suggests that indirect aerosol impacts on clouds are important but, even given the uncertainty in estimating these impacts, this mechanism is not strong enough to counter greenhouse warming effects.

Recent observational studies^{58,60} show that locally, over India, the total aerosol effect (direct plus indirect) has been associated with a surface cooling of 0.3 °C over the last three decades. This is close to the warming expected from greenhouse gases. However, the aerosols are observed to be associated with warming in the lower to middle troposphere—the regions inhabited by the glacier fields. In this case the aerosols may be enhancing the direct temperature forcing by contributing to the melting of the higher glaciers of the HKH region. **Snowfall amounts.** Aerosols are found to alter cloud physics in a manner that reduces precipitation downstream from the pollution source^{61,62}. This also reduces the snow particle rime growth, resulting in lower snow water equivalent, a result obtained from direct field measurements^{62–64}. Properly represented aerosols in climate models will apparently also work together with increasing temperature to reduce snow/ice in regions where heavy air pollution exists (for example, China, the western USA and Europe).

Snow/ice melt rates. A common aerosol found in the atmosphere over many regions of the earth is black carbon. This substance absorbs sunlight. It is scrubbed from the atmosphere by precipitation and, because it is ubiquitous, is likely to end up in the snow and ice fields of the planet. There it could decrease the surface albedo, causing the snow/ice to absorb solar energy more readily and thereby melt sooner. Measurements of black carbon amounts and its budgets are only now being made. By whatever means, darkening the surface of a snow/ice field will enhance melt rates. Again, it seems that proper inclusion of aerosols in global climate models will increase early melting of snow packs and, especially, glaciers and sea ice⁶⁵.

The bottom line here is that other important, but poorly represented, atmospheric physical and chemical processes seem unlikely to neutralize or reverse greenhouse warming. This is true even if we take the lower end of the estimated warming by the IPCC (1.4 °C) to be the net thermal forcing on the snow/glacier packs. Our ACPI study² showed that such an increase, coupled with inadequate containment, is all it takes to invoke the water storage problems noted above.

Overview of expected regional water impacts

In this review, we suggest that the simplest of changes associated with global warming (a modest increase in near-surface air temperature) will be responsible for alterations of the hydrological cycle in snowmelt-dominated regions via seasonal shifts in stream-flow. Without adequate water storage capacity, these changes will lead to regional water shortages. The model-predicted changes are already being seen in the observed data. If maintained at current levels, these changes will lead to a serious reduction in dry-season water availability in many regions of the Earth within the next few decades.

The physical principles found to apply in snowmelt-dominated regions (for example, the western USA) are one of the probable causes of the observed early snowmelt and, more importantly, deglaciation that is now occurring in most mountainous regions of the world. The serious situations developing in the HKH region and South America have been briefly presented. It is clear that both regions, as well as others not mentioned, are headed for a water-supply crisis. Better water management techniques can help, but cannot solve the problem without significant changes to agriculture, industry and lifestyle. Detailed studies of the future impact of global warming on water resources in these regions are long overdue.

a 1978



b 2002



Figure 3 | Changes in the Qori Kalis Glacier, Quelccaya Ice Cap, Peru, between 1978 (a) and 2002 (b). Glacier retreat during this time was 1,100 m (L. Thompson, personal communication). Photographs courtesy of L. Thompson.

We have discussed briefly here some of the major uncertainties in the models, in particular the impacts of aerosols and clouds, as well as their suspected impacts on the aspects of the hydrological cycle having to do with snow and ice. In all the cases considered, current scientific evidence suggests that these processes, which are currently either not included, or are marginally included, in IPCC scenario runs, will act to increase the impact of mere temperature increase on the snow and ice fields of the planet.

Time is running out for nations in the sensitive areas we have evaluated, particularly those whose water supplies are dependent on mid-latitude glaciers, to understand just what the future might hold for them. How much they can do is uncertain given the several decades of warming that will occur as a result of past actions, even if greenhouse emissions were halted at today's levels⁶⁶, but perhaps the initiation of strategic planning will be motivated by the prospect (and what is rapidly becoming the reality) of diminished water supplies.

1. The International Panel for Climate Change (IPCC) *Climate Change 2001: The Scientific Basis* (eds Houghton, J. T. et al.) (Cambridge Univ. Press, Cambridge, UK, 2001).
2. Barnett, T. P. & Pennell, W. (eds) Impact of global warming on Western US water supplies. *Clim. Change* 62 (Spec. Vol.) (2004).
3. Mote, P. W., Hamlet, A. F., Clark, M. P. & Lettenmaier, D. P. Declining mountain snow pack in western North America. *Bull. Am. Met. Soc.* 86, 39–49 (2005).
4. Dettinger, M. D., Cayan, D. R., Meyer, M. K. & Jeton, A. E. Simulated hydrologic responses to climate variations and change in the Merced, Carson, and American River Basins, Sierra Nevada, California, 1900–2099. *Clim. Change* 62, 283–317 (2004).
5. Hamlet, A. F., Mote, P. W., Clark, M. P. & Lettenmaier, D. P. Effects of temperature and precipitation variability on snow pack trends in the western U.S. *J. Clim.* (in the press).
6. Douville, H. et al. Sensitivity of the hydrological cycle in increasing amounts of greenhouse gases and aerosols. *Clim. Dyn.* 20, 45–68 (2002).
7. Giorgi, F., Whetton, P. H. & Jones, R. G. Emerging patterns of simulated regional climatic changes for the 21st century due to anthropogenic forcings. *Geophys. Res. Lett.* 28, 3317–3321 (2001).
8. Giorgi, F. & Bi, X. Regional changes in surface climate interannual variability for the 21st century from ensembles of global model simulations. *Geophys. Res. Lett.* 32, L13701, doi:10.1029/2005GL023002 (2005).
9. Ruiz-Barradas, A. & Nigam, S. IPCC's 20th century climate simulations: Varied representations of North American hydroclimate variability. *J. Clim.* (submitted).
10. Dai, A. Precipitation characteristics of eighteen coupled models. *J. Clim.* (submitted).
11. Cayan, D. R., Kammerdiener, S. A., Dettinger, M. D., Caprio, J. M. & Peterson, D. H. Changes in the onset of Spring in the Western United States. *Bull. Am. Met. Soc.* 82, 399–415 (2001).
12. Stewart, I., Cayan, D. C. & Dettinger, M. D. Changes in snowmelt runoff timing in Western North America under a 'business as usual' climate change scenario. *Clim. Change* 62, 217–232 (2004).
13. Nijssen, B., O'Donnell, G. M., Hamlet, A. F. & Lettenmaier, D. P. Hydrologic vulnerability of global rivers to climate change. *Clim. Change* 50, 143–175 (2001).
14. Liang, X., Lettenmaier, D. P., Wood, E. F. & Burges, S. J. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.* 99(D17), 14415–14428 (1994).
15. Vörösmarty, C. J. K. et al. The storage and aging of continental runoff in large reservoir systems of the world. *Ambio* 26, 210–219 (1997).
16. Vörösmarty, C. J. et al. Anthropogenic sediment retention: Major global impact from registered river impoundments. *Glob. Planet. Change* 39, 169–190 (2003).
17. Adam, J. C., Clark, E. A., Lettenmaier, D. P. & Wood, E. F. Correction of global precipitation products for orographic effects. *J. Clim.* (in the press).
18. Center for International Earth Science Information Network (CIESIN). *Socioeconomic Data and Applications Center (SEDAC): Gridded Population of the World, Version 3* (Columbia University and Centro Internacional de Agricultura Tropical, Palisades, New York, 2004); available at (<http://beta.sedac.ciesin.columbia.edu/gpw>).
19. Peterson, T. C., Golubev, V. S. & Groisman, P. V. Evaporation losing its strength. *Nature* 377, 687–688 (1995).
20. Chattopadhyay, N. & Hulme, M. Evaporation and potential evapotranspiration in India under conditions of recent and future climate change. *Agric. Forest Meteorol.* 87, 55–73 (1997).
21. Thomas, A. Spatial and temporal characteristics of potential evapotranspiration trends over China. *Int. J. Clim.* 20, 381–396 (2000).
22. Golubev, V. S. et al. Evaporation changes over the contiguous United States and the former USSR: a reassessment. *Geophys. Res. Lett.* 28(13), 2665–2668 (2001).
23. Brutsaert, W. & Parlange, M. Hydrologic cycle explains the evaporation paradox. *Nature* 396, 30 (1998).
24. Lawrimore, J. H. & Peterson, T. C. Pan evaporation trends in dry and humid regions of the United States. *J. Hydrometeorol.* 1, 543–546 (2000).
25. Hobbins, M. T. & Ramirez, J. A. Trends in pan evaporation and actual evapotranspiration across the conterminous U.S.: paradoxical or complementary? *Geophys. Res. Lett.* 31, doi: 10.1029/2004GL019846 (2004).
26. Walter, M. T., Wilks, D. S., Parlange, J.-Y. & Schneider, R. L. Increasing evapotranspiration from the conterminous United States. *J. Hydrometeorol.* 5, 405–408 (2004).
27. Roderick, M. L. & Farquhar, G. D. The cause of decreased pan evaporation over the past 50 years. *Science* 298, 1410–1411 (2002).
28. Ohmura, A. & Wild, M. Is the hydrological cycle accelerating? *Science* 298, 1345–1346 (2002).
29. Wild, M., Ohmura, A. & Gilgen, H. On the consistency of trends in radiation and temperature records and implications for the global hydrological cycle. *Geophys. Res. Lett.* 31, doi: 10.1029/2003GL019188 (2004).
30. International Panel for Climate Change. *Climate Change 2001: Impacts, Adaptation and Vulnerability* (eds McCarthy, J. J. et al.) (Cambridge Univ. Press, Cambridge, UK, 2001).
31. Barnett, T. P. et al. The effects of climate change on water resources in the West: Introduction and overview. *Clim. Change* 62, 1–11 (2004).
32. Payne, J. T., Wood, A. W., Hamlet, A. F., Palmer, R. N. & Lettenmaier, D. P. Mitigating effects of climate change on the water resources of the Columbia River Basin. *Clim. Change* 62, 233–256 (2004).
33. Middelkoop, H. et al. Impact of climate change on hydrological regimes and water resources management in the Rhine basin. *Clim. Change* 49, 105–128 (2001).
34. Gan, T. Y. Reducing vulnerability of water resources of Canadian prairies to potential droughts and possible climatic warming. *Wat. Res. Manag.* 14, 111–135 (2000).
35. Burn, D. Hydrologic effects of climatic change in west-central Canada. *J. Hydrol.* 160, 53–70 (1994).
36. de Loë, R., Kreutzweiser, R. & Moraru, L. Adaptation options for the near term: climate change and the Canadian water sector. *Glob. Environ. Change* 11, 231–245 (2001).
37. Schwindler, D. W. The cumulative effects of climate warming and other human stresses on Canadian freshwaters in the new millennium. *Can. J. Fish. Aquat. Sci.* 58, 18–29 (2001).
38. Thompson, L. G. et al. Tropical glacier and ice core evidence of climate change on annual to millennial time scales. *Clim. Change* 59, 137–155 (2003).
39. Combes, S., Prentice, M. L., Hansen, L. & Rosentrater, L. *Going, Going, Gone! Climate Change and Global Glacier Decline* 1–6 (World Wildlife Fund Climate Change Programme, WWF Germany, Berlin, 2004).
40. Singh, P. & Bengtsson, L. Hydrological sensitivity of a large Himalayan basin to climate change. *Hydrol. Process.* 18, 2363–2385 (2004).
41. Singh, P., Jain, S. K. & Kumar, N. Estimation of snow and glacier-melt contribution to the Chenab River, Western Himalaya. *Mount. Res. Develop.* 17(1), 49–56 (1997).
42. Singh, P. & Jain, S. K. Snow and glacier melt in the Satluj River at Bhakdra Dam in the western Himalayan region. *Hydrol. Sci. J.* 47, 93–106 (2002).
43. Gao, Q. & Shi, S. Water resources in the arid zone of northwest China. *J. Desert Res.* 12(4), 1–12 (1992).
44. Hou, S. et al. Climatological significance of an ice core net-accumulation record at Mt. Qomolangma. *Chin. Sci. Bull.* 45, 256–261 (2000).
45. Chinese Academy of Sciences. *China Glacier Inventory* (World Data Center for Glaciology and Geocryology, Lanzhou Institute of Glaciology and Geocryology, Lanzhou, 2004); available from NSIDC User Services (nsidc@nsidc.org).
46. Meier, M. & Dyurgerov, M. Deciphering complex changes in snow and ice. *Science* 297, 350–351 (2002).
47. Singh, P. Effect of warmer climate on the depletion of snow covered area in the Satluj basin in the western Himalayan region. *Hydrol. Sci. J.* 48, 413–425 (2003).
48. Singh, P. & Kumar, N. Impact assessment of climate change on the hydrological response of a snow and glacier melt runoff dominated Himalayan river. *J. Hydrol.* 193, 316–350 (1997).
49. Liniger, H., Weingarten, R. & Grosjean, M. *Mountains of the World: Water Towers for the 21st Century* 1–24 (Mountain Agenda, Center for Development and Environment, Institute of Geography, University of Bern, Bern, 1998).
50. Mark, B. G. & Seltzer, G. O. Tropical glacier melt water contribution to stream discharge: a case study in the Cordillera Blanca, Peru. *J. Glaciol.* 49, 271–281 (2003).
51. Kaser, G., Georges, C., Juen, I. & Moelg, T. in *Global Change and Mountain Regions: A State of Knowledge Overview* (eds Huber, U. M., Bugmann, H. K. M. & Reasoner, M. A.) 185–196 (Springer, New York, 2005).
52. Mark, B. G. & Seltzer, G. O. Evaluation of recent glacier recession in the Cordillera Blanca, Peru (AD 1962–1999): spatial distribution of mass loss and climatic forcing. *Quat. Sci. Rev.* (in the press).
53. Francou, B., Vuille, M., Wagnon, P., Mendoza, J. & Sicart, J. E. Tropical climate change recorded by a glacier in the central Andes during the last decades of the 20th century: Chacaltaya, Bolivia. *J. Geophys. Res.* 108, D54154, doi:10.1029/2002JD002959 (2003).

54. Mark, B.G. & Seltzer, G.O. in *Global Change and Mountain Regions: A State Of Knowledge Overview* (eds Huber, U. M., Bugmann, H. K. M. & Reasoner, M. A.) 205–214 (Springer, New York, 2005).
55. Vuille, M. & Bradley, R. S. Mean annual temperature trends and their vertical structure in the tropical Andes. *Geophys. Res. Lett.* **27**, 3885–3888 (2000).
56. Vuille, M., Bradley, R. S., Werner, M. & Keimig, F. 20th century climate change in the tropical Andes: observations and model results. *Clim. Change* **59**(1–2), 75–99 (2003).
57. Kaufman, Y. J., Didier, T. & Olivier, B. A satellite view of aerosols in the climate system. *Nature* **419**, 215–223 (2002).
58. Ramanathan, V., Crutzen, P. J., Kiehl, J. T. & Rosenfeld, D. Aerosols, climate, and the hydrological cycle. *Science* **294**, 2119–2124 (2001).
59. Kiehl, J. T., Schneider, T. L., Rasch, P. J. & Barth, M. C. Radiative forcing due to sulfate aerosols from simulations with the National Center for Atmospheric Research Community Climate Model, Version 3. *J. Geophys. Res.* **105**, 1441–1457 (2000).
60. Krishnan, R. & Ramanathan, V. Evidence of surface cooling from absorbing aerosols. *Geophys. Res. Lett.* **29**, 54–56 (2002).
61. Rosenfeld, D. TRMM observed first direct evidence of smoke from forest fires inhibiting rainfall. *Geophys. Res. Lett.* **26**, 3105–3108 (1999).
62. Rosenfeld, D. Suppression of rain and snow by urban air pollution. *Science* **287**, 1793–1796 (2000).
63. Borys, R. D., Lowenthal, D. H., Cohn, S. A. & Brown, W. O. J. Mountaintop and radar measurements of anthropogenic aerosol effects on snow growth and snowfall rate. *Geophys. Res. Lett.* **30**(10), 45-1–45-4 (2003).
64. Givati, A. & Rosenfeld, D. Quantifying precipitation suppression due to air pollution. *J. Appl. Met.* **43**, 1038–1056 (2004).
65. Hansen, J. & Nazarenko, L. Soot climate forcing via snow and ice albedos. *Proc. Natl Acad. Sci. USA* **101**, 423–428 (2004).
66. Hansen, J. *et al.* Earth's energy imbalance: Confirmation and implications. *Science* **308**, 1431–1435 (2005).

Acknowledgements This work is a contribution from IDAG, the International Detection and Attribution Group jointly supported by NOAA and DOE. The gross domestic product data set was developed by the Center for International Earth Science Information Network (CIESIN) at Columbia University, New York, with funding from the National Aeronautics and Space Administration. This manuscript was improved considerably through the suggestions of D. Pierce and A. Gershunov.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to T.P.B. (timdotbarnett@ucsd.edu) or D.P.L. (dennisl@u.washington.edu).

REVIEWS

Impact of regional climate change on human health

Jonathan A. Patz^{1,2}, Diarmid Campbell-Lendrum³, Tracey Holloway¹ & Jonathan A. Foley¹

The World Health Organisation estimates that the warming and precipitation trends due to anthropogenic climate change of the past 30 years already claim over 150,000 lives annually. Many prevalent human diseases are linked to climate fluctuations, from cardiovascular mortality and respiratory illnesses due to heatwaves, to altered transmission of infectious diseases and malnutrition from crop failures. Uncertainty remains in attributing the expansion or resurgence of diseases to climate change, owing to lack of long-term, high-quality data sets as well as the large influence of socio-economic factors and changes in immunity and drug resistance. Here we review the growing evidence that climate-health relationships pose increasing health risks under future projections of climate change and that the warming trend over recent decades has already contributed to increased morbidity and mortality in many regions of the world. Potentially vulnerable regions include the temperate latitudes, which are projected to warm disproportionately, the regions around the Pacific and Indian oceans that are currently subjected to large rainfall variability due to the El Niño/Southern Oscillation sub-Saharan Africa and sprawling cities where the urban heat island effect could intensify extreme climatic events.

Global average temperatures are projected to increase between 1.4 and 5.8 °C by the end of this century¹; an associated rise in sea level is also expected. The number of people at risk from flooding by coastal storm surges is projected to increase from the current 75 million to 200 million in a scenario of mid-range climate changes, in which a rise in the sea level of 40 cm is envisaged by the 2080s (ref. 2). Extremes of the hydrologic cycle (such as floods and droughts) are projected to increase with warmer ambient temperatures. Evidence is mounting that such changes in the broad-scale climate system may already be affecting human health, including mortality and morbidity from extreme heat, cold, drought or storms; changes in air and water quality; and changes in the ecology of infectious diseases^{3–5}.

We reviewed both empirical studies of past observations of climate-health relationships, and model simulation studies of projected health risks and regional vulnerability associated with future climate change. Here we focus on the health implications of climate variability, past and present climate change impacts on human health, future projections and uncertainties. This review primarily examines relatively direct-acting temperature effects, while recognizing that other major risk pathways exist, for instance, altered storm patterns, hydrologic extremes, and sea-level rise.

Health implications of climate variability

Non-infectious health effects. The summer of 2003 was probably Europe's hottest summer in over 500 years, with average temperatures 3.5 °C above normal^{6–8}. With approximately 22,000 to 45,000 heat-related deaths occurring across Europe over two weeks in August 2003 (refs 9 and 10), this is the most striking recent example of health risks directly resulting from temperature change. Judging from this extreme event, changes in climate variability associated with long-term climate change could be at least as important for future risk assessment as upward trends in mean temperature.

The European heatwave in 2003 was well outside the range of expected climate variability⁸. In addition, comparisons of climate model outputs with and without anthropogenic drivers show that the risk of a heatwave of that magnitude had more than doubled by 2003 as a result of human-induced climate change³. The demonstration of a causal link between global warming and the occurrence of regional heatwaves indicates a potential for more frequent and/or more severe heatwaves in a future warmer world.

On local and regional scales, changes in land cover can sometimes exacerbate the effect of greenhouse-gas-induced warming, or even exert the largest impact on climatic conditions. For example, urban 'heat islands' result from lowered evaporative cooling, increased heat storage and sensible heat flux caused by the lowered vegetation cover, increased impervious cover and complex surfaces of the cityscape. Dark surfaces such as asphalt roads or rooftops can reach temperatures 30–40 °C higher than surrounding air¹¹. Most cities show a large heat island effect, registering 5–11 °C warmer than surrounding rural areas¹². But the effects of land cover change on climate are not limited to small areas: at the scale of the entire continental USA, Kalnay and Cai¹³ estimated that land-cover changes (from both agriculture and urban areas) caused a surface warming of ~0.27 °C per century. Also, in southeast China, a warming of ~0.05 °C per decade since 1978 has been attributed to land-use change from urban sprawl¹⁴.

Exposure to both extreme hot and cold weather is associated with increased morbidity and mortality, compared to an intermediate 'comfortable' temperature range¹⁵. Heat mortality follows a J-shaped function with a steeper slope at higher temperatures¹⁶. The comfortable or safest temperature range is closely related to mean temperature, with an upper bound from as low as 16.5 °C for the Netherlands and 19 °C for London¹⁷, to as high as 29 °C in Taiwan¹⁸. Hot days occurring earlier in the summer season have a larger effect than those occurring later¹⁷. It should be noted that although the majority of temperature-mortality studies have taken place in developed

¹Center for Sustainability and the Global Environment (SAGE), Nelson Institute for Environmental Studies, and ²the Department of Population Health Sciences, University of Wisconsin, 1710 University Avenue, Madison, Wisconsin 53726, USA. ³Department of Protection of the Human Environment, World Health Organization, Geneva, Avenue Appia, Geneva CH-1211, Switzerland.

countries and in regions with temperate climates, the same pattern of temperature–mortality relationship found in European and North American cities occurs in São Paulo, Brazil, a developing city with subtropical conditions¹⁹.

In summary, although most studies to date show clear vulnerability to heat in cooler temperate regions, tropical regions may well show a similar sensitivity as location-specific temperatures rise.

Climatic influences on regional famines are another well-recognized climate–health association. Malnutrition remains one

of the largest health crises worldwide and according to the WHO, approximately 800 million people are currently undernourished, with close to half of these living in Africa²⁰. Droughts and other climate extremes have direct impacts on food crops, and can also influence food supply indirectly by altering the ecology of plant pathogens. Projections of the effect of climate change on food crop yield production globally appear to be broadly neutral, but climate change will probably exacerbate regional food supply inequalities²¹.

Infectious diseases. Climatic variations and extreme weather events have profound impacts on infectious disease. Infectious agents (such as protozoa, bacteria and viruses) and their associated vector organisms (such as mosquitoes, ticks and sandflies) are devoid of thermodynamic mechanisms, and reproduction and survival rates are thus strongly affected by fluctuations in temperature^{4,22}. Temperature dependencies are seen in correlations between disease rates and weather variations over weeks, months or years²³ and in close geographic associations between key climate variables and the distributions of important vector-borne diseases^{24,25}.

Malaria transmission has been associated with anomalies of maximum temperature in the highlands of Kenya²⁶. Several studies of long-term trends in malaria incidence and climate in Africa, however, have not found a link to temperature trends, emphasizing instead the importance of including other key determinants of malaria risk such as drug resistance, human migration and immune status, inconsistent vector- or disease-control programmes, and local land-use changes^{27–30}. However, in the highland Debre Zeit sector of central Ethiopia an association has been documented between increasing malaria prevalence and incidence with concomitant warming trends from 1968 to 1993 (ref. 31). Controlling for confounding factors, the association could not be explained by drug resistance, population migration, or level of vector-control efforts. In short, studies of the association of malaria and past climate in the African Highlands remains controversial in part due to varying quality of long-term disease data across sites in Africa, and in part due to the difficulty in adequately controlling for demographic and biological (drug resistance) data. A definitive role of long-term climate trends has not been ascertained.

Dengue fever and the more serious form of this disease, dengue haemorrhagic fever (DHF), are caused by the world's most prevalent mosquito-borne virus. All strains of the dengue virus are carried principally by the *Aedes aegypti* mosquito. This mosquito is strongly affected by ecological and human drivers, particularly the density of water-bearing containers, but is also influenced by climate, including variability in temperature, moisture and solar radiation. For relatively small countries with presumably some climate uniformity, a climate-based dengue model has been developed that strongly correlates with the inter-annual variability in dengue cases reported at the national level (Fig. 1)³².

A few examples of other vector-borne diseases demonstrating variance with climate include the Ross River virus in Australia^{33,34}, and plague³⁵ in the American southwest. Bluetongue, a disease of livestock, has increased its northern range in Europe since 1998, paralleling trends in warming and controlling for many biological and socioeconomic factors³⁶.

Temperature has also been found to affect food-borne infectious diseases. For example, higher than average temperatures contribute to an estimated 30% of reported cases of salmonellosis across much of continental Europe³⁷. In the UK, the monthly incidence of food poisoning is most strongly associated with the temperatures occurring in the previous two to five weeks³⁸.

El Niño/Southern Oscillation and infectious diseases. With the exception of seasonal variability, the El Niño/Southern Oscillation (ENSO) is the strongest naturally occurring source of climate variability around the globe³⁹. Studies of malaria have revealed the health impacts of interannual climate variability associated with El Niño, including large epidemics on the Indian subcontinent⁴⁰, in Colombia⁴¹, Venezuela⁴² and Uganda⁴³. Rift Valley fever epidemics

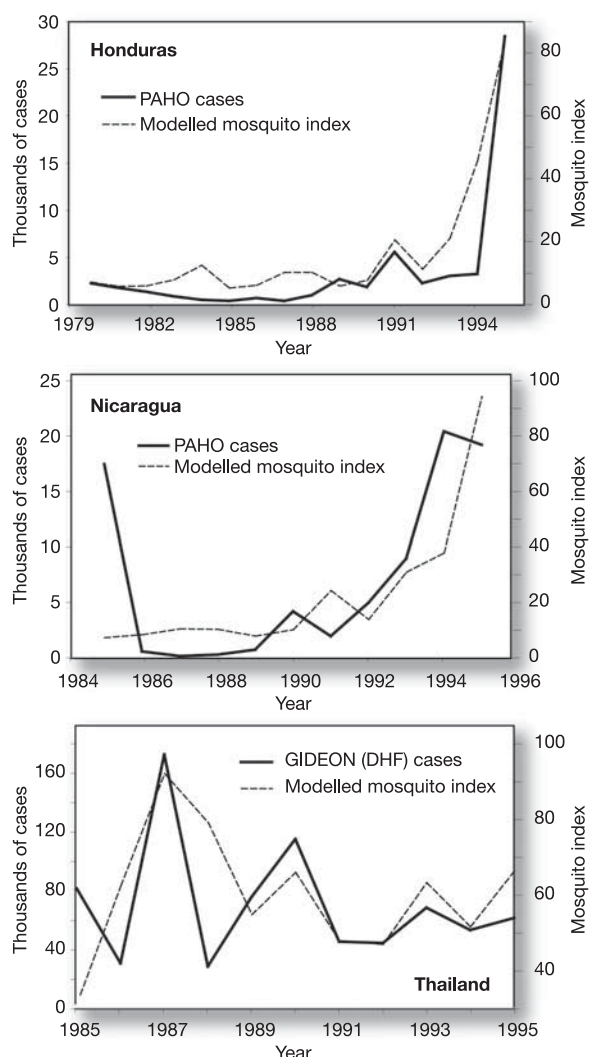


Figure 1 | Correlation between simulated, climate-driven variations in *Aedes aegypti* mosquito density and observed variations in dengue and DHF cases. Using a computer model of mosquito physiology and development, estimated changes in the relative abundance of *Aedes aegypti* that were driven only by month-to-month and year-to-year variations in temperature, humidity, solar radiation and rainfall were analysed. The simulated, climate-induced variations in mosquito density were then compared to reported cases of dengue and DHF across many nations of the world that covered at least one degree of latitude and longitude and had at least five years of dengue caseload data. In many countries of Central America and Southeast Asia, the relationship is statistically significant ($P < 0.05$). For example, climate-driven fluctuations in *Ae. aegypti* densities appear to be related to annual variations in dengue/DHF cases in Honduras, Nicaragua and Thailand as shown. These represent relatively small-area countries; for larger countries endemic for dengue such as Brazil, China, India and Mexico, the association is not significant, as might be expected because the disease data was at the country level. Graphs adapted from ref. 32.

between 1950 and 1998 have coincided with unusually high rainfall in East Africa associated with ENSO-related Pacific and Indian Ocean sea surface temperature (SST) anomalies⁴⁴. While more than three quarters of the Rift Valley Fever outbreaks between 1950 and 1988 occurred during warm ENSO event periods⁴⁵, some epidemics have also occurred in years with no El Niño, and the model has not been validated against new epidemics.

A 'wavelet analysis' method was recently used to incorporate host immunity and pathogen population dynamics of DHF in Thailand.

A spatial-temporal travelling wave explained a three-year period cycle in disease incidence, starting in Bangkok, moving radially at a speed of 148 km per month⁴⁶. In a subsequent study that controlled for this intrinsic synchronization, El Niño remained as a significant determinant of dengue epidemics that cycled every two to three years from 1986 to 1992 in Thailand⁴⁷.

Hantavirus pulmonary syndrome in the American southwest can be predicted on the basis of ENSO events; following the 1991–92 El Niño, associated heavy rainfall led to an increase in the

Table 1 | Global burden of climate-change-attributable disease

Region	CVD		Diarrhoea			Malaria			Floods			
	Mortality*	Risk‡	Mortality*	Disease†	Risk‡	Mortality*	Disease†	Risk‡	Mortality*	Disease†	Risk‡	
											Inland	Coastal
AFR-D	1	1.007	5	154	1.08	5	178	1.02	0	1	1.36	1.64
AFR-E	1	1.005	8	260	1.08	18	682	1.14	0	3	1.48	1.18
AMR-A	0	1	0	0	1	0	0	1.51	0	4	4.93	1.19
AMR-B	1	1.004	0	0	1	0	3	1.15	1	67	2.13	2.27
AMR-D	0	1.005	1	17	1.02	0	0	1.08	0	5	1.78	4.64
EMR-B	0	1.003	0	14	1	0	0	1	0	6	2.67	1.75
EMR-D	1	1.003	8	277	1.09	3	112	1.29	1	46	3.05	3.91
EUR-A	0	0.999	0	0	1	0	0	1	0	3	3.55	1.14
EUR-B	0	0.999	0	6	1.01	0	0	1	0	4	1.82	6.31
EUR-C	0	0.998	0	3	1	0	0	1.48	0	1	2.35	1.04
SEAR-B	1	1.007	1	28	1	0	0	1	0	6	1.79	1.39
SEAR-D	7	1.007	22	612	1.09	0	0	1.01	0	8	1.12	1.04
WPR-A	0	0.999	0	0	1	0	0	1.48	0	1	1.76	1.04
WPR-B	0	1	2	89	1.01	1	43	1.42	0	37	1.62	1.05
World	12§	–	47	1,459	–	27	1,018	–	2	193	–	–

Region	Malnutrition			All causes		Total deaths per million	Total DALYs per million
	Mortality*	Disease†	Risk‡	Mortality*	Disease†		
AFR-D	8	293	1.02	19	626	66.83	2,185.78
AFR-E	9	323	1.02	36	1,267	109.4	3,839.58
AMR-A	0	0	1	0	4	0.15	11.85
AMR-B	0	0	1	2	71	3.74	166.62
AMR-D	0	0	1	1	23	10.28	324.15
EMR-B	0	0	1	1	20	5.65	147.57
EMR-D	9	313	1.08	21	748	61.3	2,145.91
EUR-A	0	0	1	0	3	0.07	6.66
EUR-B	0	0	1	0	10	1.04	48.13
EUR-C	0	0	1	0	4	0.29	14.93
SEAR-B	0	0	1	2	34	7.91	117.19
SEAR-D	52	1,918	1.17	80	2,538	65.79	2,080.84
WPR-A	0	0	1	0	1	0.09	8.69
WPR-B	0	0	0.99	3	169	2.16	111.36
World	77	2,846	–	166	5,517	27.82	925.35

* Estimated mortality in thousands attributable to climate change in 2000 (compared to baseline climate of 1961–1990).

† Estimated disease burden in thousands of DALYs attributable to climate change in 2000.

‡ Projected changes in relative risk for 2030.

§ Heat-related deaths without subtracting potential reductions in cold-related deaths; this value was therefore not included in the aggregate estimates of mortality due to climate change.

The data in Table 1 are taken from ref. 57. The region key is taken from ref. 57. AFR-D: Algeria, Angola, Benin, Burkina Faso, Cameroon, Cape Verde, Chad, Comoros, Equatorial Guinea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Madagascar, Mali, Mauritania, Mauritius, Niger, Nigeria, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Togo.

AFR-E: Botswana, Burundi, Central African Republic, Congo, Côte d'Ivoire, Democratic Republic of the Congo, Eritrea, Ethiopia, Kenya, Lesotho, Malawi, Mozambique, Namibia, Rwanda, South Africa, Swaziland, Uganda, United Republic of Tanzania, Zambia, Zimbabwe.

AMR-A: Cuba, Canada, United States of America.

AMR-B: Antigua and Barbuda, Argentina, Bahamas, Barbados, Belize, Brazil, Chile, Colombia, Costa Rica, Dominica, Dominican Republic, El Salvador, Grenada, Guyana, Honduras, Jamaica, Mexico, Panama, Paraguay, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Suriname, Trinidad and Tobago, Uruguay, Venezuela.

AMR-D: Bolivia, Ecuador, Guatemala, Haiti, Nicaragua, Peru.

EMR-B: Bahrain, Cyprus, Iran, Jordan, Kuwait, Lebanon, Libyan Arab Jamahiriya, Oman, Qatar, Saudi Arabia, Syrian Arab Republic, Tunisia, United Arab Emirates.

EMR-D: Afghanistan, Djibouti, Egypt, Iraq, Morocco, Pakistan, Somalia, Sudan, Yemen.

EUR-A: Andorra, Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, Malta, Monaco, the Netherlands, Norway, Portugal, San Marino, Slovenia, Spain, Sweden, Switzerland, United Kingdom.

EUR-B: Albania, Armenia, Azerbaijan, Bosnia and Herzegovina, Bulgaria, Georgia, Kyrgyzstan, Poland, Romania, Slovakia, Tajikistan, Macedonia, Turkey, Turkmenistan, Uzbekistan, Yugoslavia.

EUR-C: Belarus, Estonia, Hungary, Kazakhstan, Latvia, Lithuania, Moldova, Russian Federation, Ukraine.

SEAR-B: Indonesia, Sri Lanka, Thailand.

SEAR-D: Bangladesh, Bhutan (Democratic People's Republic of), Korea, India, Maldives, Myanmar, Nepal.

WPR-A: Australia, Brunei, Darussalam, Japan, New Zealand, Singapore.

WPR-B: Albania, China, Cook Islands, Fiji, Kiribati, Lao, Malaysia, Marshall Islands, Micronesia, Mongolia, Nauru, Niue, Palau, Papua New Guinea, Philippines, Republic of Korea, Samoa, Solomon Islands, Tonga, Tuvalu, Vanuatu, Vietnam.

rodent populations that preceded human cases of disease⁴⁸. Based on these climate/ecology/disease relationships, a climate- and GIS (Geographic Information System)-based model was developed that predicted disease risk reasonably well for the following strong El Niño event of 1997–98 (ref. 49).

Waterborne diseases, such as childhood diarrhoeal disease, are also influenced by El Niño, as was observed with the 1997–98 El Niño event in Peru. During that unseasonable winter, the ambient temperature in Lima increased more than 5 °C above normal, and the number of daily admissions for diarrhoea increased by more than twofold, compared to expected trends⁵⁰.

Cholera has varied with climatic fluctuations and SSTs affected by the ENSO phenomenon over multi-decadal time periods in Bangladesh⁵¹. In the Bay of Bengal, upward trends in cholera also have been linked to longer-term climate changes (that is, changes over approximately a century), with weak cholera/ENSO links found during 1893–1940, and strong and consistent associations occurring during the more pronounced ENSO fluctuations between 1980–2001 (ref. 52). One ecologically based hypothesis for this link involves copepods (zooplankton), which feed on algae, and can serve as reservoirs for *Vibrio cholerae* and other enteric pathogens⁵³; copepods bloom in response to the warming SSTs generally associated with El Niño.

Understanding interannual cycles of cholera and other infectious diseases (as seen above for dengue fever), however, requires the combined analyses of both environmental exposures and intrinsic host immunity to a disease. When these factors are considered together, interannual variability of cholera is strongly correlated to SSTs in the Bay of Bengal, ENSO, the extent of flooding in Bangladesh across short time periods (<7 years), and to monsoon rains and Brahmaputra river discharge for longer period climate patterns (>7 years)⁵⁴.

Although it is not clear whether and how ENSO dynamics will change in a warmer world, regions that are currently strongly affected by ENSO (for example, southeast Asia, southern and east Africa, the southwest USA, and various regions of South America) could experience heightened risks if ENSO variability, or the strength of El Niño events intensifies.

Land use, local climate and infectious disease. Just as the ‘urban heat island’ (mentioned above) exacerbates heatwaves, so too can land use change influence transmission of infectious diseases. Land

cover may affect mosquito habitat by changing local temperature and humidity. For example, temperatures were significantly higher in communities in highland Uganda bordering cultivated swamps compared with natural ones⁵⁵, and average minimum temperatures were associated with the number of *Anopheles gambiae sensu lato* (s.l.) mosquitoes per house after adjustment for potential confounding variables. Of course, in some locations, rainfall is more important than temperatures or land use, and in some locations malaria epidemics in the highlands of Africa are more influenced by changing disease-control efforts than any other factors²⁹.

Past and present climate-change impacts

In the most comprehensive, peer-reviewed and quantitative climate-health assessment to date, the World Health Organization (WHO) examined the global burden of disease already attributable to anthropogenic climate change up to the year 2000 (ref. 20); WHO also made model-based forecasts of the health risks from global climate change until 2030 (refs 56, 57).

The study made generally conservative assumptions about climate-health relationships (for example, that socioeconomic conditions would prevent a climate-driven spread of vector-borne disease from endemic tropical regions to temperate regions), and health impacts were included only if quantitative models were available. An assessment over such a broad range of health impacts is by nature approximate, as there are significant uncertainties in all climate change–disease models. The study indicates that the climatic changes that have occurred since the mid-1970s could already be causing over 150,000 deaths and approximately five million ‘disability-adjusted life years’ (DALYs) per year through increasing incidences of diseases such as diarrhoea (temperature effects only), malaria and malnutrition that occur mainly in developing countries⁵⁷ (Table 1, Fig. 2).

The WHO assessment emphasized that actions to adapt to a changing climate will require regional assessments of vulnerability to specific health risks, and interventions that are geographically and temporally targeted on highly susceptible populations.

Future projections and uncertainties

The WHO extended its estimates of morbidity and mortality caused by human-induced climate change to the year 2030, following

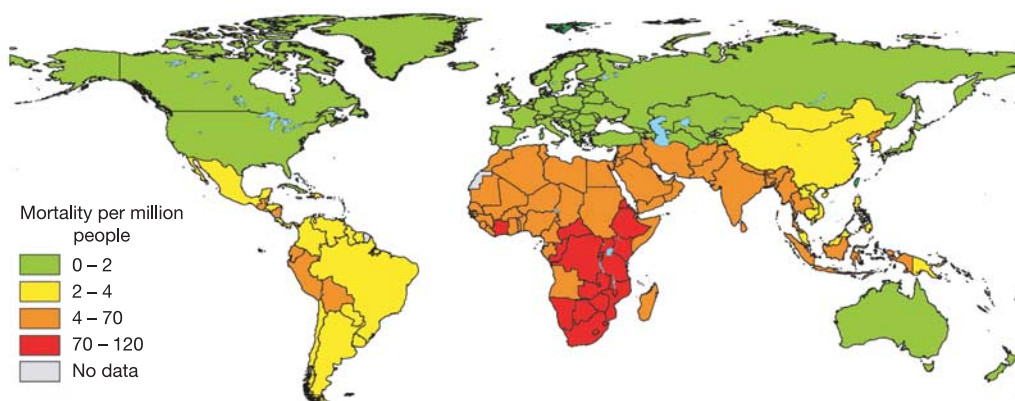


Figure 2 | WHO estimated mortality (per million people) attributable to climate change by the year 2000. The IPCC ‘business as usual’ greenhouse gas emissions scenario, ‘IS92a’ and the HadCM2 GCM of the UK Hadley Centre were used to estimate climate changes relative to ‘baseline’ 1961–1990 levels of greenhouse gases and associated climate conditions. Existing quantitative studies of climate–health relationships were used to estimate relative changes in a range of climate-sensitive health outcomes including: cardiovascular diseases, diarrhoea, malaria, inland and coastal

flooding, and malnutrition, for the years 2000 to 2030. This is only a partial list of potential health outcomes, and there are significant uncertainties in all of the underlying models. These estimates should therefore be considered as a conservative, approximate, estimate of the health burden of climate change. Even so, the total mortality due to anthropogenic climate change by 2000 is estimated to be at least 150,000 people per year. Details on the methodology are contained in ref. 57.

Hadley Centre global climate model (GCM) projections for a range of greenhouse-gas emissions scenarios⁵⁸. It estimates that the climate-change-induced excess risk of the various health outcomes will more than double by the year 2030 (ref. 57). Large increases are predicted for the relative risk of flooding and more modest changes in diseases such as malaria, malnutrition and diarrhoea (Table 1). However, it is important to note that these small relative changes may actually cause far greater aggregate disease burdens. In sub-Saharan Africa, for example, flooding currently kills less than one person per million annually, while malaria kills over 1,600 per million and diarrhoea kills over 1,000 per million (ref. 57).

To consider changes in future heatwave probabilities, GCM projections of future climate for conditions contributing to heatwaves are now capable of estimating the occurrence of stagnant, warm air masses that can determine the severity of a heatwave, including variables such as consecutive nights (three or more) with high minimum night-time temperatures⁵⁹. A recent analysis of the 1995 Chicago and 2003 Europe heatwaves predicted intensified magnitude and duration of heatwaves over portions of Europe and the United States, suggesting that heatwaves in Chicago and Paris will be 25% and 31% more frequent, respectively, by 2090 and that the average length of a heatwave in Paris will have increased from 8–13 days to 11–17 days. Large increases in heatwaves were also projected for the western and southern USA and the Mediterranean region⁵⁹.

Data from the MARA (Mapping Malaria Risk in Africa) project have been applied to global climate projections to examine potential changes in malaria risk over regions of Africa⁶⁰. Excluding any increase in population, an increase of 16–28% in person-month exposure (number of people exposed per month) to malaria risk by year 2100 was determined⁶⁰. However, like all previous continental or global models of malaria–climate relationships, the study fails to account for non-climatic determinants or the variation of specific climate–disease relationships among locations^{61,62}.

Extrapolation from statistically based models into the future is of limited value. There may be evidence that malaria is increasing in the highlands of Africa owing to climate change, but methods used to detect it are still controversial and do not convincingly prove or disprove the association. To assess the health risks of mid- to long-term future climate projections, a concerted effort combining the use

of process-based models (capturing the biology of the malaria system) alongside statistical modelling will be needed.

Regional assessments of health impacts. Climate-change projections from GCMs, such as a recent set of simulations performed in preparation for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report⁶³, are increasing in resolution, but are still not appropriate for analysing disease patterns at scales smaller than a few grid boxes of area 250 km². For example, local patterns of climate that are strongly influenced by subtle changes in topography (slope and aspect, rainshadow effects, orographic precipitation), large water bodies, coastlines and other geographic features may be important determinants of disease ecology.

Moving from large-scale climate projections to smaller spatial scales requires the application of ‘downscaling’ techniques that bring additional information to bear on the region in question. Downscaling methods fall into two broad categories: dynamical downscaling, using high-resolution, regional climate models^{64,65}, and statistical downscaling, based on statistical relationships between large-scale predictor variables and regional predictants (Table 2).

A dynamical downscaling study with the aim of determining the impact of potential climate changes over the next 50 years on air pollution in the eastern USA⁶⁶ reported that under the high-emission ‘A2’ IPCC scenarios, daily average ozone levels increase by 3.7 p.p.b. across the eastern USA, with the most polluted cities today experiencing the greatest increase in temperature-related ozone pollution (Fig. 3). Across 15 selected cities in this region, the average number of days exceeding the 8-hour ozone standard increased by 60%—from 12 to almost 20 days per summer by the 2050s (ref. 67). Assuming constant population and dose–response characteristics, an independent dynamical downscaling study⁶⁴ (refs 64 and 67 both stem from the modeling work of ref. 66) projected that ozone-related deaths from climate change will increase by ~4.5% for the mid-2050s (using the ‘A2’ emissions scenario), compared with the levels of the 1990s. Considering the potential population exposed to outdoor air pollution (in the millions), this seemingly small relative risk actually translates to quite a substantial attributable health risk. There is significant uncertainty associated with these findings, as they are based on a single emissions scenario, one GCM simulation, and many assumptions about regional ozone precursor emissions.

Table 2 | Differences between dynamical and statistical downscaling

	Benefits	Drawbacks	Applications
Dynamical downscaling	• Simulates climate mechanisms • No <i>a priori</i> assumptions about how current and future climate are related • ‘State of the science’ tools • Continually advancing computers are making RCMs faster and cheaper to run • Encourages collaborations between health and climate scientists	• Expensive, in terms of computer resources and professional expertise • Results may be sensitive to uncertain parameterisations • Biases in the GCM (providing boundary conditions) may propagate to regional scale • Output from models may not be in a format well-suited to health analysis—additional data processing often required	• Health responses associated with climate extremes and nonlinear variability • Data-poor areas • Connecting outcomes with climate processes • Include land-use impacts on climate or health outcomes
Statistical downscaling (especially regression methods)	• Much cheaper (runs quickly on desktop computers with free software) • Builds on the statistical expertise common among public health researchers • May correct for biases in GCM • Allows for the assessment of climate results over a range of GCMs and emission scenarios	• Assumes relationships between local and large-scale climate remain constant • Does not capture climate mechanisms • Not well suited to capturing variance or extreme events	• Climate means, and variability with some limitations • Data-rich regions, especially Northern Hemisphere mid-latitudes • Compare present with projected climate in a consistent framework • Test a range of inputs • Variable scales, down to individual measurement sites

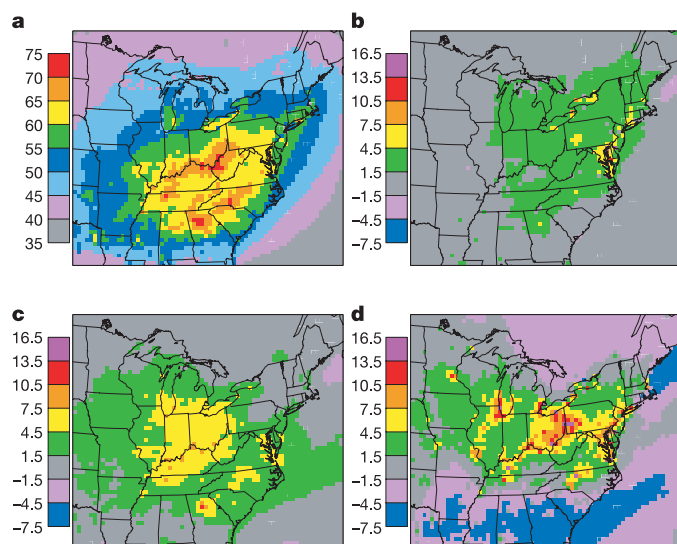


Figure 3 | Simulated ozone air pollution over the eastern United States by using a downscaled climate model linked to a regional air pollution model. a, Baseline summertime average daily maximum 8-hour O_3 concentrations for the 1990s. **b–d**, The following panels show changes in summertime-average daily maximum 8-hour O_3 concentrations for the 2020s (**b**), the 2050s (**c**), and the 2080s (**d**) over the region based on IPCC A2 scenario simulations relative to the 1990s, in parts per billion. Five consecutive summer seasons were simulated in each decade starting with the NASA Goddard Institute for Space Studies (GISS) Atmosphere–Ocean Global Climate Model, with results subsequently downscaled using the mesoscale regional climate model (MM5), and finally coupled to the Community Multiscale Air Quality (CMAQ) model. Simulation results for the 2020s, 2050s and 2080s indicate that summertime average daily maximum 8-hour O_3 concentrations increase by 2.7, 4.2 and 5.0 p.p.b., respectively, as a result of regional climate change. The data were taken from ref. 66.

Statistical downscaling is most useful for health assessment in data-rich regions. On the basis of statistical downscaling techniques, heat-related deaths in California are estimated to more than double by the year 2100 (ref. 68).

Pursuing early warning systems. Current early warning systems for infectious diseases are beginning to show some utility. For example, over two-thirds of the inter-annual variability of malaria in Botswana can be predicted from the SSTs and associated monthly rainfall⁶⁹. For more direct health impacts from heatwaves, although uncertainty still remains as to which weather parameters are most hazardous, a number of studies have consistently identified high minimum (night-time) temperatures, duration, and early seasonal occurrence of heatwaves as particularly dangerous conditions⁷⁰. Therefore, early warning systems may offer some health protection from the effects of heatwaves. For example, after the 1995 heatwave in the United States, the city of Milwaukee initiated an ‘extreme heat conditions plan’ involving local agencies, communications tests, stepped responses to early forecasts, a 24-hour ‘hotline’ and other interventions. Reductions in heat-related morbidity (measured by emergency ambulance runs) and mortality were reduced by 49% from expected levels during a heatwave in 1999, and were not attributable to differences in heat levels alone⁷¹.

Currently, over two dozen cities worldwide have a ‘synoptic-based’ weather-watch warning system, which focuses monitoring on dangerous air masses⁷². These systems successfully forecast most days with excess deaths in Rome during the 2003 heatwave⁷³, and have been implemented successfully in Shanghai, for example⁷⁴. However, variability in predictability between cities suggests that systems must be location-specific, requiring the input of considerable amounts of health-related and meteorological data for each locale⁷⁵.

Health and regional climate change

Two main climatic impacts on health at a regional scale emerge from this review: direct heat-related mortality and morbidity, and a climate-mediated change in the incidence of infectious diseases.

Heat-related mortality is dominated by the difference between temperature extremes and the mean climate—especially early in summer when people have not yet become accustomed to higher temperatures—rather than by gradual increases in mean temperatures. Projections of future climate suggest such increases in extremes in relation to mean temperatures may occur particularly in the mid-latitudes. In addition, the effect of heatwaves is exacerbated in large cities owing to the urban heat island effect. As urban areas and urban population grow, vulnerability to heat-related mortality seems likely to increase in the future.

Studies of climatic influences on infectious diseases have mainly focused on the influence of ENSO. ENSO has been found to be related to incidences of malaria in South America, rift valley fever in east Africa, dengue fever in Thailand, hantavirus pulmonary syndrome in the southwestern USA, childhood diarrhoeal disease in Peru and cholera in Bangladesh. It is unclear at this stage whether global warming will significantly increase the amplitude of ENSO variability, but if so, the regions surrounding the Pacific and Indian oceans are expected to be most vulnerable to the associated changes in health risks.

Potential impacts of long-term trends in mean temperatures on health, for example, on malaria incidence in the African highlands, have not been reliably detected. The data available at present do not allow robust control for non-climatic confounding factors such as socio-economic influences, immunity patterns and drug resistance effects. However, regions bordering areas with high endemicity of climate-sensitive diseases, where temperatures at present limit the geographic distribution of disease (such as malaria in the African highlands) could be at risk in a warmer climate.

Early warning systems both for heatwaves and for expected outbreaks of infectious diseases can help to adapt to some of the effects of a changing climate, through measures such as opening air-conditioned shopping malls at night-time to those who are most vulnerable to heat, or providing prophylactic treatment to those in danger from infectious diseases. However, population vulnerability still greatly depends on economic and other determinants of a society’s capacity to provide such measures.

Land use and land cover change, as mentioned above, can magnify the effects of extreme climatic events, both on direct health outcomes (for example, heat mortality), and on ecologically mediated infectious diseases in any region of the world. Therefore, to assess accurately future climate-change impacts on health, future projections of land-use change must be considered as well.

As illustrated in Fig. 2, the regions with the greatest burden of climate-sensitive diseases are also the regions with the lowest capacity to adapt to the new risks. Africa—the continent where an estimated 90% of malaria occurs—has some of the lowest per capita emissions of the greenhouse gases that cause global warming. In this sense, global climate change not only presents new region-specific health risks, but also a global ethical challenge. To meet this challenge, precautionary approaches to mitigating anthropogenic greenhouse gases will be necessary, while research continues on the full range of climate–health mechanisms and potential future health impacts.

1. Intergovernmental Panel on Climate Change. *Climate Change 2001: The Scientific Basis: Contribution of Working Group I to the Third Assessment Report* 1–944 (Cambridge Univ. Press, Cambridge, UK, 2001).
2. Intergovernmental Panel on Climate Change. *Climate Change 2001: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Third Assessment Report* 1–1000 (Cambridge Univ. Press, Cambridge, UK, 2001).
3. Stott, P. A., Stone, D. A. & Allen, M. R. Human contribution to the European heatwave of 2003. *Nature* **432**, 610–614 (2004).
4. Kovats, R. S., Campbell-Lendrum, D. H., McMichael, A. J., Woodward, A. & Cox, J. S. Early effects of climate change: Do they include changes in vector-borne disease? *Phil. Trans. R. Soc. Ser. B* **356**, 1057–1068 (2001).

5. Patz, J. A., Epstein, P. R., Burke, T. A. & Balbus, J. M. Global climate change and emerging infectious diseases. *J. Am. Med. Assoc.* **275**, 217–223 (1996).
6. Beniston, M. The 2003 heatwave in Europe: A shape of things to come? An analysis based on Swiss climatological data and model simulations. *Geophys. Res. Lett.* **31**, 2022–2026 (2004).
7. Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M. & Wanner, H. European seasonal and annual temperature variability, trends, and extremes since 1500. *Science* **303**, 1499–1503 (2004).
8. Schar, C. *et al.* The role of increasing temperature variability in European summer heatwaves. *Nature* **427**, 332–336 (2004).
9. International Federation of Red Cross and Red Crescent Societies. *World Disaster Report 2004* Ch. 2 (IFRC, 2004).
10. Kosatsky, T. The 2003 European heat waves. *Euro Surveill.* Published online July 2005, 10(7), (<http://www.eurosurveillance.org/em/v10n07/1007-222.asp>).
11. Frumkin, H. Urban sprawl and public health. *Public Health Rep.* **117**, 201–217 (2002).
12. Aniello, C., Morgan, K., Busbey, A. & Newland, L. Mapping micro-urban heat islands using Landsat Tm and a GIS. *Comput. Geosci.* **21**, 965–969 (1995).
13. Kalnay, E. & Cai, M. Impact of urbanization and land-use change on climate. *Nature* **423**, 528–531 (2003).
14. Zhou, L. M. *et al.* Evidence for a significant urbanization effect on climate in China. *Proc. Natl Acad. Sci. USA* **101**, 9540–9544 (2004).
15. Curriero, F. C., Heiner, K., Zeger, S., Samet, J. M. & Patz, J. A. Temperature and mortality in 11 cities of the eastern United States. *Am. J. Epidemiol.* **155**, 80–87 (2002).
16. McMichael, A. J., Haines, A., Slooff, R. & Kovats, S. (eds) *Climate Change and Human Health* (The World Health Organization, Geneva, 1996).
17. Hajat, S., Kovats, R. S., Atkinson, R. W. & Haines, A. Impact of hot temperatures on death in London: a time series approach. *J. Epidemiol. Commun. Health* **56**, 367–372 (2002).
18. Martens, W. J. Health impacts of climate change and ozone depletion: an ecoepidemiologic modeling approach. *Environ. Health Perspect.* **106** (Suppl 1), 241–251 (1998).
19. Gouveia, N., Hajat, S. & Armstrong, B. Socioeconomic differentials in the temperature-mortality relationship in São Paulo, Brazil. *Int. J. Epidemiol.* **32**, 390–397 (2003).
20. The World Health Organisation. *The World Health Report 2002* (WHO, Geneva, 2002).
21. Parry, M. L., Rosenzweig, C., Iglesias, A., Livermore, M. & Fischer, G. Effects of climate change on global food production under SRES emissions and socio-economic scenarios. *Glob. Environ. Change* **14**, 53–67 (2004).
22. Gubler, D. J. *et al.* Climate variability and change in the United States: Potential impacts on vector- and rodent-borne diseases. *Environ. Health Perspect.* **109**, 223–233 (2001).
23. Kuhn, K. G., Campbell-Lendrum, D. H., Armstrong, B. & Davies, C. R. Malaria in Britain: past, present, and future. *Proc. Natl Acad. Sci. USA* **100**, 9997–10001 (2003).
24. Hales, S., de Wet, N., Maindonald, J. & Woodward, A. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *Lancet* **360**, 830–834 (2002).
25. Rogers, D. J. & Randolph, S. E. The global spread of malaria in a future, warmer world. *Science* **289**, 1763–1766 (2000).
26. Githeko, A. K. & Ndegwa, W. Predicting malaria epidemics in the Kenyan Highlands using climate data: a tool for decision makers. *Glob. Change Hum. Health* **2**, 54–63 (2001).
27. Hay, S. I. *et al.* Climate change and the resurgence of malaria in the East African highlands. *Nature* **415**, 905–909 (2002).
28. Mouchet, J. Malaria epidemics on the Highlands of Madagascar and of East and South Africa. *Bull. Soc. Pathol. Exot.* **91**, 64–66 (1998).
29. Mouchet, J. *et al.* Evolution of malaria in Africa for the past 40 years: impact of climatic and human factors. *J. Am. Mosq. Control Assoc.* **14**, 121–130 (1998).
30. Shanks, G. D., Hay, S. I., Stern, D. I., Biomndo, K. & Snow, R. W. Meteorologic influences on *Plasmodium falciparum* malaria in the highland tea estates of Kericho, western Kenya. *Emerg. Infect. Dis.* **8**, 1404–1408 (2002).
31. Tulu, A. N. *Determinants of Malaria Transmission in the Highlands of Ethiopia. The Impact of Global Warming on Morbidity and Mortality Ascribed to Malaria.* PhD thesis, Univ. London (1996).
32. Hopp, M. J. & Foley, J. A. Worldwide fluctuations in dengue fever cases related to climate variability. *Clim. Res.* **25**, 85–94 (2003).
33. Tong, S. L., Hu, W. B. & McMichael, A. J. Climate variability and Ross River virus transmission in Townsville region, Australia, 1985–1996. *Trop. Med. Int. Health* **9**, 298–304 (2004).
34. Woodruff, R., Guest, C., Garner, G., Becker, N. & Lindsay, M. F. Weather and climate as early warning system indicators for epidemics of Ross River virus: a case study in south-west western Australia. *Epidemiology* **14**, S94–S94 (2003).
35. Parmenter, R. R., Yadav, E. P., Parmenter, C. A., Ettestad, P. & Gage, K. L. Incidence of plague associated with increased winter-spring precipitation in New Mexico. *Am. J. Trop. Med. Hyg.* **61**, 814–821 (1999).
36. Purse, B. V. *et al.* Climate change and the recent emergence of bluetongue in Europe. *Nature Rev. Microbiol.* **3**, 171–181 (2005).
37. Kovats, R. S. *et al.* The effect of temperature on food poisoning: a time-series analysis of salmonellosis in ten European countries. *Epidemiol. Infect.* **132**, 443–453 (2004).
38. Bentham, G. & Langford, I. H. Environmental temperatures and the incidence of food poisoning in England and Wales. *Int. J. Biometeorol.* **45**, 22–26 (2001).
39. Ropelewski, C. F. & Halpert, M. S. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. *Mon. Weath. Rev.* **111**, 517–528 (1987).
40. Bouma, M. J. & van der Kaay, H. J. The El Niño Southern Oscillation and the historic malaria epidemics on the Indian subcontinent and Sri Lanka: an early warning system for future epidemics? *Trop. Med. Int. Health* **1**, 86–96 (1996).
41. Poveda, G. *et al.* Coupling between annual and ENSO timescales in the malaria-climate association in Colombia. *Environ. Health Perspect.* **109**, 489–493 (2001).
42. Bouma, M. J. & Dye, C. Cycles of malaria associated with El Niño in Venezuela. *J. Am. Med. Assoc.* **278**, 1772–1774 (1997).
43. Lindblade, K. A., Walker, E. D., Onapa, A. W., Katungu, J. & Wilson, M. L. Highland malaria in Uganda: prospective analysis of an epidemic associated with El Niño. *Trans. R. Soc. Trop. Med. Hyg.* **93**, 480–487 (1999).
44. Linthicum, K. J. *et al.* Climate and satellite indicators to forecast Rift Valley fever epidemics in Kenya. *Science* **285**, 397–400 (1999).
45. Anyamba, A., Linthicum, K. J. & Tucker, C. J. Climate-disease connections: Rift Valley Fever in Kenya. *Cad Saude Publica* **17** (Suppl.), 133–140 (2001).
46. Cummings, D. A. *et al.* Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. *Nature* **427**, 344–347 (2004).
47. Czelles, B., Chavez, M., McMichael, A. J. & Hales, S. Nonstationary influence of El Niño on the synchronous dengue epidemics in Thailand. *PLoS Med.* **2**, e106 (2005).
48. Glass, G. E. *et al.* Using remotely sensed data to identify areas at risk for hantavirus pulmonary syndrome. *Emerg. Infect. Dis.* **6**, 238–247 (2000).
49. Glass, G. E. *et al.* Satellite imagery characterizes local animal reservoir populations of Sin Nombre virus in the southwestern United States. *Proc. Natl Acad. Sci. USA* **99**, 16817–16822 (2002).
50. Checkley, W. *et al.* Effect of El Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children. *Lancet* **355**, 442–450 (2000).
51. Pascual, M., Rodo, X., Ellner, S. P., Colwell, R. & Bouma, M. J. Cholera dynamics and El Niño-Southern Oscillation. *Science* **289**, 1766–1769 (2000).
52. Rodo, X., Pascual, M., Fuchs, G. & Faruque, A. S. ENSO and cholera: a non-stationary link related to climate change? *Proc. Natl Acad. Sci.* **99**, 12901–12906 (2002).
53. Colwell, R. R. Global climate and infectious disease: the cholera paradigm. *Science* **274**, 2025–2031 (1996).
54. Koelle, K., Rodo, X., Pascual, M., Yunus, M. & Mostafa, G. Refractory periods and climate forcing in cholera dynamics. *Nature* **436**, 696–700 (2005).
55. Lindblade, K. A., Walker, E. D., Onapa, A. W., Katungu, J. & Wilson, M. L. Land use change alters malaria transmission parameters by modifying temperature in a highland area of Uganda. *Trop. Med. Int. Health* **5**, 263–274 (2000).
56. McMichael, A. J. Impact of climatic and other environmental changes on food production and population health in the coming decades. *Proc. Nutr. Soc.* **60**, 195–201 (2001).
57. McMichael, A. J. *et al.* in *Comparative Quantification of Health Risks: Global and Regional Burden of Disease due to Selected Major Risk Factors* (eds Ezzati, M., Lopez, A. D., Rodgers, A. & Murray, C. J. L.) Ch. 20, 1543–1649 (World Health Organization, Geneva, 2004).
58. Arnell, N. W. The consequences of CO₂ stabilization for the impacts of climate change. *Clim. Change* **53**, 413–446 (2002).
59. Meehl, G. A. & Tebaldi, C. More intense, more frequent, and longer lasting heatwaves in the 21st century. *Science* **305**, 994–997 (2004).
60. Tanser, F. C., Sharp, B. & le Sueur, D. Potential effect of climate change on malaria transmission in Africa. *Lancet* **362**, 1792–1798 (2003).
61. Hay, S. I. *et al.* Climate variability and malaria epidemics in the highlands of East Africa. *Trends Parasitol.* **21**, 53–63 (2005).
62. Reiter, P. *et al.* Global warming and malaria: a call for accuracy. *Lancet Infect. Dis.* **4**, 323–324 (2004).
63. Meehl, G. A., Covey, C., McAvaney, B., Latif, M. & Stouffer, R. J. Overview of the Coupled Model Intercomparison Project. *Bull. Am. Meteorol. Soc.* **86**, 89–93 (2005).
64. Knowlton, K. *et al.* Assessing ozone-related health impacts under a changing climate. *Environ. Health Perspect.* **112**, 1557–1563 (2004).
65. McMichael, A. *et al.* *Human Health and Climate Change in Oceania: A Risk Assessment* (Commonwealth of Australia, Canberra, 2003).
66. Hogrefe, C. *et al.* Simulating changes in regional air pollution over the eastern United States due to changes in global and regional climate and emissions. *J. Geophys. Res.* **109**, 2627–2638 (2004).
67. Patz, J. A. *et al.* *Heat Advisory: Climate Change, Air Pollution, and Health in the US* (Natural Resources Defense Council, Washington, 2004).
68. Hayhoe, K. *et al.* Emissions pathways, climate change, and impacts on California. *Proc. Natl Acad. Sci. USA* **101**, 12422–12427 (2004).
69. Thomson, M. C., Mason, S. J., Phindela, T. & Connor, S. J. Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana. *Am. J. Trop. Med. Hyg.* **73**, 214–221 (2005).

70. Sheridan, S. C. & Kalkstein, L. S. Health watch/warning systems in urban areas. *World Res. Rev.* **10**, 375–383 (1998).
71. Weisskopf, M. G. *et al.* Heatwave morbidity and mortality, Milwaukee, Wis, 1999 vs 1995: an improved response? *Am. J. Public Health* **92**, 830–833 (2002).
72. Sheridan, S. C. & Kalkstein, L. S. Progress in heat watch-warning system technology. *Bull. Am. Meteorol. Soc.* **85**, 1931–1941 (2004).
73. Michelozzi, P. *et al.* Impact of heatwaves on mortality—Rome, Italy, June–August 2003. *J. Am. Med. Assoc.* **291**, 2537–2538 (2004); reprinted from *Morb. Mort. Weekly Rep.* **53**, 369–371 (2004).
74. Tan, J. G. *et al.* An operational heat/health warning system in Shanghai. *Int. J. Biometeorol.* **48**, 157–162 (2004).
75. Ebi, K. L., Teisberg, T. J., Kalkstein, L. S., Robinson, L. & Weiher, R. F. Heat watch/warning systems save lives—Estimated costs and benefits for Philadelphia 1995–98. *Bull. Am. Meteorol. Soc.* **85**, 1067–1073 (2004).

Acknowledgements We thank contributors to the climate change portion of the WHO Global Burden of Disease Assessment led by T. McMichael for the data and analysis incorporated into Table 1 and Fig. 2. For constructing figures and tables and assisting in manuscript preparation, we thank S. Olson, E. Sowatzke, M. Sternitzky and D. Zaks. We thank C. Tebaldi for providing technical information on the 2003 European heatwave, and C. Hogrefe for providing Fig. 3. We also thank M. Pascual and J. Ahumada for assistance with manuscript revisions. The views expressed in this article are those of the authors and do not necessarily reflect the position of the World Health Organization.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence should be addressed to J.A.P. (patz@wisc.edu).

ARTICLES

Structure of the *E. coli* protein-conducting channel bound to a translating ribosome

Kakoli Mitra^{1*}, Christiane Schaffitzel^{2*}, Tanvir Shaikh³, Florence Tama⁴, Simon Jenni², Charles L. Brooks III⁴, Nenad Ban² & Joachim Frank^{1,3,5}

Secreted and membrane proteins are translocated across or into cell membranes through a protein-conducting channel (PCC). Here we present a cryo-electron microscopy reconstruction of the *Escherichia coli* PCC, SecYEG, complexed with the ribosome and a nascent chain containing a signal anchor. This reconstruction shows a messenger RNA, three transfer RNAs, the nascent chain, and detailed features of both a translocating PCC and a second, non-translocating PCC bound to mRNA hairpins. The translocating PCC forms connections with ribosomal RNA hairpins on two sides and ribosomal proteins at the back, leaving a frontal opening. Normal mode-based flexible fitting of the archaeal SecYEG structure into the PCC electron microscopy densities favours a front-to-front arrangement of two SecYEG complexes in the PCC, and supports channel formation by the opening of two linked SecY halves during polypeptide translocation. On the basis of our observation in the translocating PCC of two segregated pores with different degrees of access to bulk lipid, we propose a model for co-translational protein translocation.

The correct functioning of every living cell requires that secreted, soluble proteins be translocated across, and membrane proteins be integrated into, a cellular membrane. Protein translocation at the membrane occurs through a proteinaceous channel¹, the translocon², at the core of which lies the PCC. The PCC is a heterotrimeric integral membrane protein complex^{3,4} composed of the α -subunit (SecY in eubacteria and archaea; Sec61 α in mammals), the β -subunit (SecE in eubacteria; Sec β in archaea; Sec61 β in mammals) and the γ -subunit (SecE in eubacteria and archaea; Sec61 γ in mammals). Most membrane proteins and some soluble proteins are translocated co-translationally, a process in which the ribosome binds to the PCC, and translocation of the nascent chain is concomitant with polypeptide elongation on the ribosome. The PCC is a dynamic complex, which must be able to open and close an aqueous channel^{5,6} that is relatively isolated from hydrophobic lipids and runs perpendicular to the membrane plane to allow hydrophilic regions of a polypeptide to cross. The PCC must also be able to open and close laterally⁷ to regulate the lipid-mediated partitioning of hydrophobic transmembrane helices (TMHs) into the plane of the lipid bilayer⁸.

Low-resolution cryo-electron microscopy (cryo-EM) studies of eukaryotic co-translational translocation complexes^{9–12} have demonstrated that the functional PCC is an oligomer of the Sec61 $\alpha\beta\gamma$ complex. The recent X-ray structure of a non-translocating, monomeric, archaeal SecYEG heterotrimer¹³—uncomplexed with a ribosome or substrate polypeptide—shows that the α -subunit, SecY, is divided into two independent amino-terminal (TMHs 1–5) and carboxy-terminal (TMHs 6–10) halves, forming a ‘clam shell’¹⁴. One side of this ‘clam shell’ is hinged by the loop between TMHs 5 and 6 and clamped together by the γ -subunit, SecE, while the other side—forming the lateral gate—is unconstrained. Each SecY half contributes to the formation of a transmembrane funnel-like cavity in the centre of the complex, which is blocked by a plug (TMH 2a). It has

been proposed that the signal sequence or TMH signal anchor of a translocating nascent chain displaces the plug and wedges itself into the lateral gate, thus opening the two SecY halves in the plane of the membrane and enabling either the translocation of a hydrophilic region of a nascent chain across the lipid bilayer, or the lateral partitioning of a hydrophobic TMH into the lipid bilayer^{13,14}. Additionally, the X-ray structure shows that the long cytoplasmic loops between TMHs 6 and 7, and TMHs 8 and 9, extend ~20 Å above the membrane plane for interaction with cytosolic factors, such as the large ribosomal subunit^{15–17}.

We have used a combination of single-particle cryo-EM and computational methods to determine and interpret the structure of a functional, co-translational translocation complex from *E. coli*, comprising of SecYEG and a ribosome–nascent chain (RNC) complex. Our cryo-EM reconstruction is of sufficient resolution to determine the precise number of SecYEG monomers in the PCC and to differentiate, using computational methods, between two principal models of SecYEG monomer arrangements, and therefore helps to clarify the structural and mechanistic details of co-translational translocation.

Overall structure of the *E. coli* RNC–SecYEG complex

E. coli ribosomes programmed in a cell-free translation system with mRNA encoding a chimaeric nascent chain (Supplementary Fig. 1), which includes the TMH signal anchor of FtsQ and the SecM stalling sequence¹⁸, were complexed with detergent-solubilized SecYEG. These reconstituted RNC–SecYEG complexes were used for structure determination by cryo-EM using single-particle reconstruction. The electron microscopy data (Fourier shell correlation characteristics ~11 Å at 3 σ , ~15 Å at 0.5) extended out to ~1/11 Å⁻¹ in Fourier space; that is, some structural features down to 11 Å in size could be discerned (Supplementary Fig. 4). In this reconstruction, tRNAs are

¹Howard Hughes Medical Institute, Health Research Inc. at the Wadsworth Center, Empire State Plaza, Albany, New York 12201-0509, USA. ²Institute for Molecular Biology and Biophysics, Eidgenössische Technische Hochschule Hönggerberg, CH-8093 Zürich, Switzerland. ³Wadsworth Center, Empire State Plaza, Albany, NY 12201-0509, USA.

⁴Department of Molecular Biology, The Scripps Research Institute, 10550 Torrey Pines Road, La Jolla, California 92037, USA. ⁵Department of Biomedical Sciences, State University of New York at Albany, Empire State Plaza, Albany, New York 12201-0509, USA.

*These authors contributed equally to this work.

visible in all three canonical sites, A, P and E (Fig. 1a). The path of over 100 nucleotides of the mRNA, from its entrance to past its exit site in the 30S subunit, can be traced, as can most of the path of the nascent chain in the polypeptide exit tunnel. Greatly improving on the globular appearance of the PCC in previous single-particle cryo-EM studies^{9–12}, detailed rod- and lamella-like features, corresponding to groupings of transmembrane helices in the PCC, are discernible in the current reconstruction (Fig. 1a–c). As observed previously, there is a large frontal opening between the ribosome and the PCC at the polypeptide exit site, through which the translocating polypeptide is accessible (Fig. 1b, and see Supplementary Discussion 2.1 for a discussion on the misinterpretation of existing fluorescence data, leading to perceived discrepancies with structural data).

The observation of a second PCC bound at the exit of the mRNA channel (occupancy ~70%) was unexpected. When viewed perpendicularly to the membrane plane (Fig. 1c, e), the shapes of the two PCCs are different: the PCC at the mRNA channel exit is an elliptic cylinder (dimensions 110 Å × 70 Å, height ~45 Å) with pseudo-222

symmetry (Supplementary Fig. 2), whereas the PCC at the polypeptide exit site comprises a circular cylinder (~95 Å diameter, height ~45 Å). Because the latter PCC contains the TMH signal anchor of FtsQ (see Methods), it represents a translocating state. The PCC at the mRNA channel exit is non-physiologically bound and so must represent a non-translocating state. Volume calculations on the isolated densities of both PCCs indicate that each PCC contains two copies of SecYEG, as also suggested biochemically^{19,20} (see Supplementary Results 1.1, Supplementary Discussion 2.2 and Supplementary Fig. 5 for a discussion on the effect of low resolution on PCC electron microscopy density).

Normal mode-based flexible fitting of PCC atomic coordinates

Two copies of a homology-based atomic model of *E. coli* SecYEG (SecYEG_{Ec}; see Methods) were docked rigidly into the isolated densities of both the non-translocating and the translocating PCC. While docking into the translocating PCC density was unsuccessful, a good fit (cross-correlation coefficient (CC) of 0.78) to the non-translocating PCC density was achieved when two SecYEG_{Ec} heterotrimers were placed in a front-to-front arrangement. Thus, the structure of the PCC at the mRNA channel exit may represent a physiologically relevant non-translocating state, providing a starting structure for determining the proper conformation of the translocating PCC at the polypeptide exit site. As the relative motions of the N- and C-terminal halves of SecY are postulated to underlie SecYEG function in a translocating PCC^{13,14}, we opted to utilize a computational method—normal mode-based flexible fitting (NMFF)²¹. Normal modes provide information on the natural vibrations of a molecule and the preferential directions of collective, many-atom displacements, while maintaining the steric and geometric constraints of the overall structure. Normal mode analysis (NMA)²² applied to plug-less SecYEM_J (see Methods) confirmed the expected opening motion of linked SecY halves¹³ (Supplementary Fig. 3). We then proceeded to use a progression of docking and NMFF steps (Supplementary Results 1.2 and Supplementary Fig. 4) to obtain a model for the translocating PCC, in both a front-to-front arrangement and the recently proposed back-to-back¹³ arrangement.

The front-to-front model fitted well (CC_{final} 0.79) into the non-translocating PCC density, while the back-to-back model fitted poorly (CC_{final} 0.64). The bulk of each dimeric PCC model could be docked into the translocating PCC density (CC_{initial} 0.65 (front-to-front), 0.61 (back-to-back)) only when the ribosome-binding cytoplasmic loops between TMHs 6 and 7 and TMHs 8 and 9 (combined cross-section of C α backbone ~20 Å) of both SecYEG heterotrimers were docked into the connection densities to the ribosome located on the sides (C1, C2), which have a roughly circular cross-section of ~20 Å (back connection (C3) cross-sectional dimensions ~45 × 20 Å) (Fig. 2b, c, f, g). On the application of NMFF to these models, the two SecY halves in each heterotrimer in the front-to-front model (CC_{final} 0.79) show marked opening motions within the membrane plane (distance between the centres of mass of linked SecY halves increases by an average of ~5 Å) (Fig. 2a, d, i, j), while no opening motion is seen in the back-to-back model (CC_{final} 0.65). Thus, the formation of a pore wide enough to enable polypeptide translocation, a process thought to be necessary for translocation^{13,14}, is not observed in the back-to-back model as in the front-to-front case (Fig. 2a, d, e, h and Supplementary Fig. 4c, d). SecYEG has been shown to be inactivated when SecE is disulphide-crosslinked to a neighbouring SecE or SecY molecule, indicating that SecE dynamics, which are restricted when SecE clamps are juxtaposed tightly as in the back-to-back arrangement, are essential for protein translocation^{23,24}.

Our experimental and computational data thus provide two main arguments—one structural, the other functional—for why the back-to-back model is less favourable. Furthermore, key biochemical observations cannot be explained on the basis of a back-to-back arrangement of SecYEG heterotrimers, in which the lateral gates in

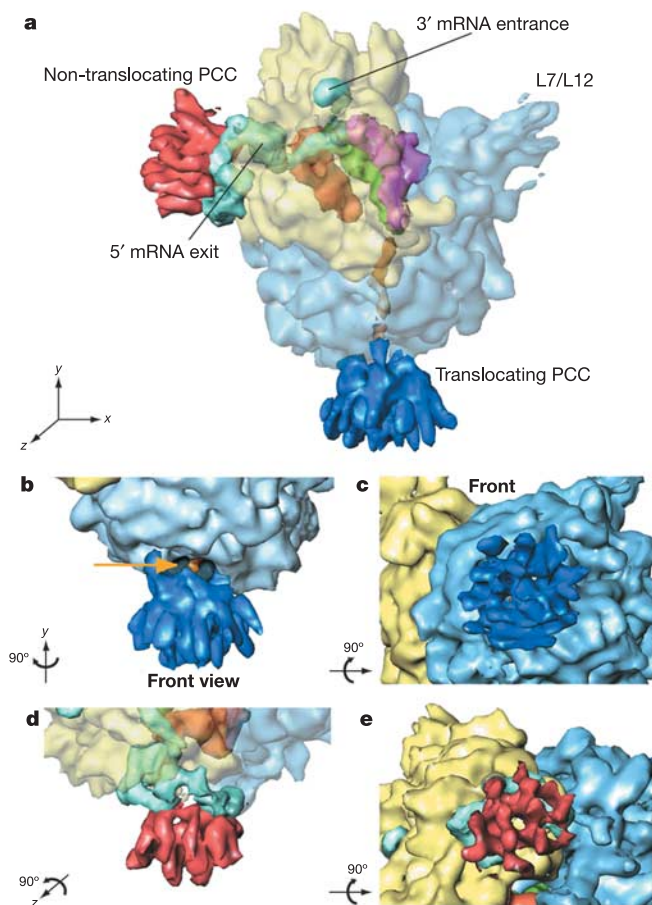


Figure 1 | General features of the cryo-EM reconstruction of the *E. coli* RNC-SecYEG complex. Cryo-EM densities are shown for the small (30S, yellow) and large (50S, sky blue) subunits, and the A-, P- and E-site tRNAs (magenta, green and orange, respectively). Isolated densities for the mRNA (cyan), the nascent chain in the ribosomal polypeptide exit tunnel (gold), the PCC at the polypeptide exit site (translocating PCC, dark blue), and the PCC at the 5' mRNA channel exit (non-translocating PCC, red) are also shown. **a**, General overview of the RNC-SecYEG complex. L7/L12, stalk of L7/L12 protein. **b**, **c**, Close-up views of the translocating PCC, showing the front view (**b**) with the nascent chain visible through the large frontal opening between the ribosome and the PCC (orange arrow), and the line of view perpendicular to the plane of the membrane (**c**). **d**, **e**, Close-up views of the non-translocating PCC.

each heterotrimer face away from each other towards the bulk lipid in the membrane. During post-translational translocation—a process in which a fully translated pre-protein is translocated through the PCC in the absence of a ribosome—disulphide-bonded, partially folded regions of the pre-protein can cross through the PCC channel²⁵. Forming an aqueous pore large enough for this process^{25,26} in the back-to-back model would require the lateral gate in a heterotrimer to open up beyond the point where the wedged signal sequence could sterically block lipid diffusion into the pore, making the translocation of a hydrophilic polypeptide energetically very unfavourable. Similarly, at a certain point during co-translational translocation, the signal sequence or signal anchor is no longer

associated with SecYEG/Sec61 $\alpha\beta\gamma$ (ref. 27), leaving the lateral gate permeable to lipids and making the continued translocation of a hydrophilic nascent chain energetically very unfavourable. However, the front-to-front model is consistent with these biochemical observations and certain observed crosslinks in SecY (ref. 20) because the lateral gates of the two heterotrimers face each other; thus, a consolidated channel—segregated from lipid—could potentially form.

Details of ribosome–PCC interactions

In order to determine the details of interaction between the PCC and ribosomal elements, an *E. coli* ribosome atomic model was refined into our cryo-EM reconstruction (see Methods). As previously observed^{11,12}, both ribosomal proteins and RNA surround the polypeptide exit site at which the PCC is located. Ribosomal proteins L17, L32 and L22 are located in the front, above—but not interacting with—the PCC (Fig. 3a), whereas proteins L24, L29 and L23 interact with the PCC in the back (Fig. 3b). The PCC makes three connections with the ribosome (Figs 2b, c and 3c). The connections at the sides, C1 and C2, are interactions between rRNA hairpins (helix 59 in C1 and helix 24 in C2) and the cytosolic factor-associating domains (CFAD) of SecY (the cytoplasmic regions between TMHs 6 and 7 and TMHs 8 and 9) (Fig. 3a, c), as has been suggested biochemically^{15–17}. The wall of connection at the back (C3) seems to be mediated mostly through interactions of L29 and L23 with the cytoplasmic region of the SecG subunit²⁸ and possibly of the N-terminal part of SecE (Fig. 3b, c). Neither the SecG/ β /61 β subunit nor the N-terminal portion of SecE are essential for viability^{29,30}. However, SecG/ β /61 β has been observed to stabilize the ribosome–PCC complex and facilitate nascent chain translocation^{29,31}. Therefore, connection C3 may lend stability and directionality (to be discussed later) to the PCC–ribosome association, whereas connections C1 and C2 may be required for PCC attachment to the ribosome¹⁶.

Modelling of 121 nucleotides of RNA into the mRNA density isolated from our reconstruction, using sequence-specific secondary structure information, indicated the presence of two major mRNA hairpins at the channel exit, immediately above the CFAD of both heterotrimers of the non-translocating PCC (Fig. 3d). Biochemical competition experiments have demonstrated that mRNA can compete with rRNA for PCC binding¹⁵. This inferred interaction of the CFADs with immobilized, structured mRNA hairpins is analogous to the interaction observed at the polypeptide exit site with rRNA hairpins, and may explain the occurrence of the non-translocating PCC at the mRNA channel exit. It is possible that the structure of this PCC may reflect that of a non-translocating PCC when it is bound to the ribosome at the polypeptide exit site.

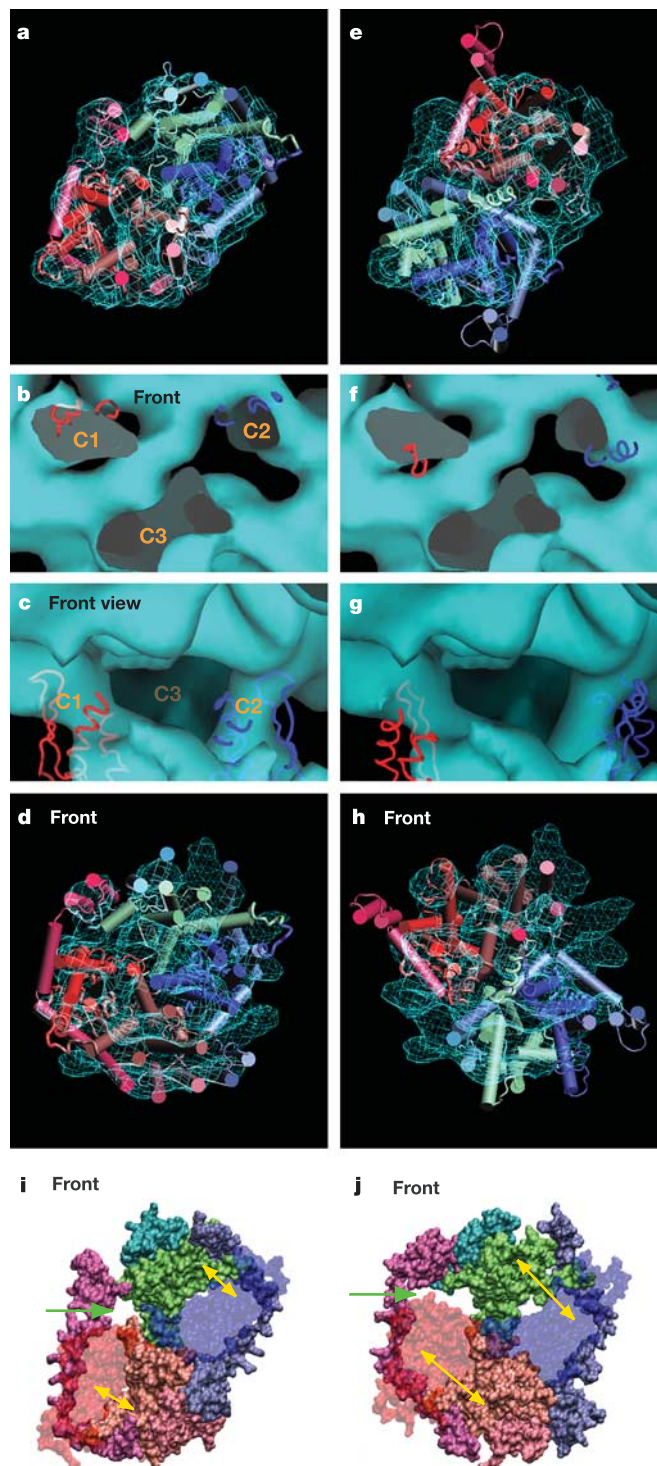


Figure 2 | Normal mode-based flexible fitting (NMFF) of the SecYEG complex into cryo-EM density. **a, d**, Model after NMFF of a front-to-front arrangement of two SecYEG_{Ec} heterotrimers into the non-translocating (**a**) and translocating (**d**) PCC density. **b, c**, Placement of the non-translocating front-to-front model into the translocating PCC density, such that the ribosome-binding SecY cytoplasmic loops are placed within connection densities. The three connection regions (C1–C3) are shown in cross-section (line of view perpendicular to the membrane plane) (**b**) and in the front view (**c**), with the frontal opening visible between C1 and C2. Electron microscopy density is shown in cyan as a semi-transparent surface. **e–h**, Parallel analysis of the back-to-back SecYEG_{Ec} model. **i, j**, van der Waals surface representations of the front-to-front models obtained by fitting to the non-translocating (**i**) and translocating (**j**) PCC density (C-terminal SecY halves are transparent for clarity). The green and yellow arrows indicate the change in the heterotrimer interface at the front, and linked SecY-half opening, respectively. In the atomic models, one heterotrimer is coloured in shades of blue/green, the other in shades of red. The view is within the plane of the membrane, with the ribosomal side behind the plane. The atomic models in **a, d, e**, and **h**, with TMHs depicted as rods, are shown docked into the experimental electron microscopy density (shown as a cyan mesh).

The path of the translocating nascent chain

On NMFF of the dimeric plug-less SecYEG complex into the translocating PCC density, prominent regions of density remain unaccounted for (green and yellow asterisks, Fig. 4a), most notably a

long rod-like region (green asterisk) at the front interface of the two heterotrimers. This density most probably corresponds to the nascent chain TMH signal anchor, close to SecY TMHs 2 and 7, as suggested by crosslinking experiments³². Furthermore, the

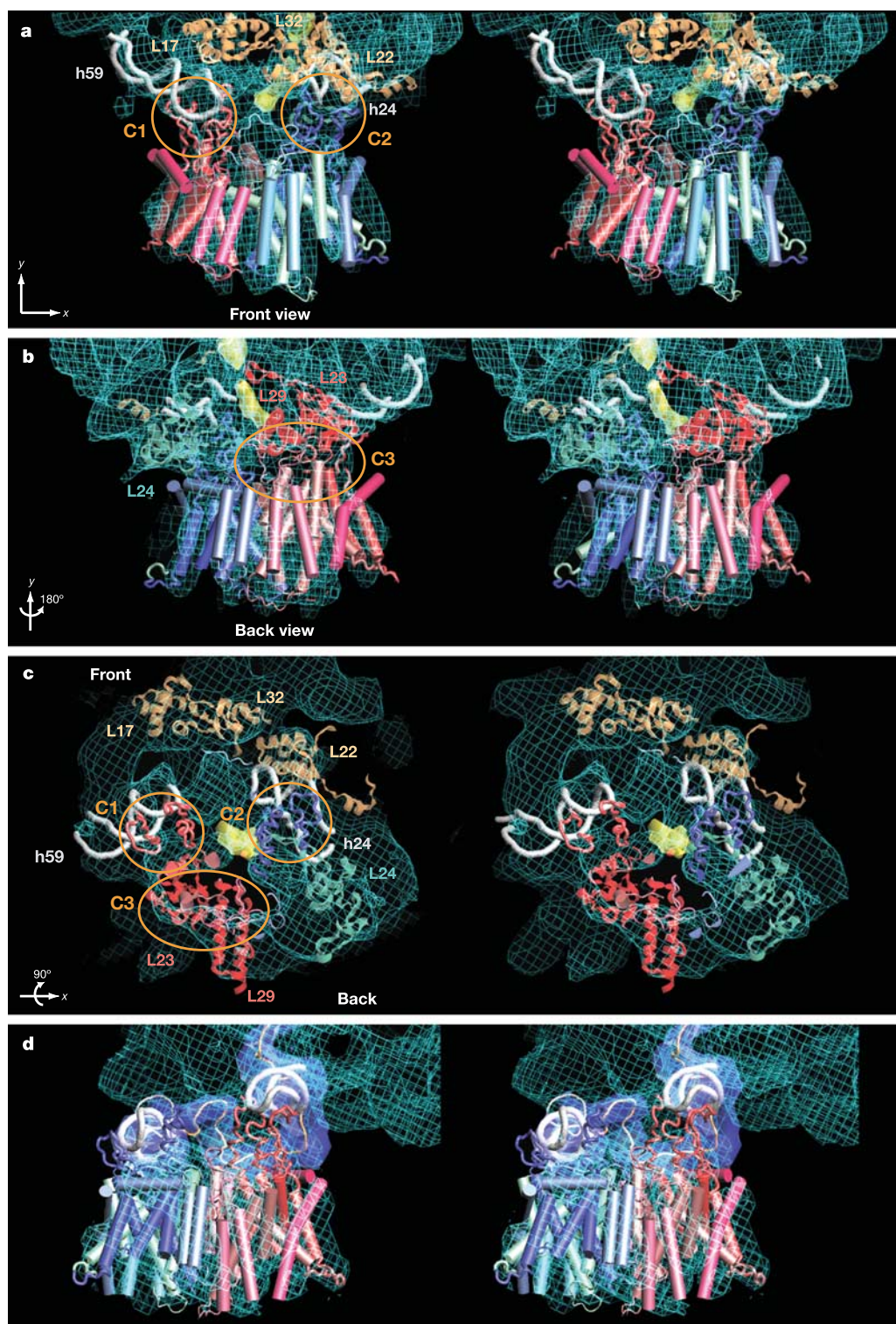


Figure 3 | Stereo views of RNA and protein elements in the ribosome-PCC junction. Real-space refined models showing *E. coli* ribosomal proteins rendered as ribbons, and rRNA regions interacting with the PCC as thick, light-grey backbone snake-like structures (rattlers). The PCC is coloured and rendered as in Fig. 2d with the cryo-EM density in cyan mesh. **a–c**, The ribosome-PCC junction at the polypeptide exit site (nascent chain density semi-transparent yellow surface) is shown in the front (**a**) and back (**b**)

views. Ribosomal elements near the polypeptide exit site and parts of the PCC in the three connection regions are also illustrated (**c**), with the line of view perpendicular to the membrane plane. Connection regions between ribosome and PCC are circled in orange and labelled. **d**, The non-translocating PCC uses its CFADs to interact with hairpins in the mRNA (mRNA shown as semi-transparent purple surface with non-interacting mRNA rattler regions in yellow). h, helix.

observation that the interface formed by the two heterotrimers at the front changes markedly from the non-translocating to the translocating state may reflect signal anchor binding at this site (Fig. 2i, j). The other empty regions of density (yellow asterisks) in the middle of both heterotrimers could represent either the plug (TMH 2a) or the remaining, hydrophilic region of the nascent chain. An atomic model was fitted into the isolated thread of density corresponding to the nascent chain inside the ribosome, which facilitated the identification of putative interactions with ribosomal RNA (K.M., C.S., F. Fabiola, M. S. Chapman, N.B. and J.F., manuscript submitted) and protein. The real-space refined positions of ribosomal protein domains suggest an interaction of the nascent chain with finger-

like projections of proteins L4 and L22 in the polypeptide exit tunnel, as has been demonstrated biochemically^{18,33}. Towards the end of the polypeptide exit tunnel, the projection of L23 is in proximity to the nascent chain, indicating another potential interaction site (Fig. 4b, d).

Co-translational translocation through the PCC

The cryo-EM reconstruction of the *E. coli* RNC–SecYEG presented in this study offers an insight into the architecture of the ribosome-bound PCC in its translocating, and possibly non-translocating, state. Analysis of atomic models obtained by NMFF suggests that the PCC is most probably formed by a front-to-front association of two SecYEG/Sec61 $\alpha\beta\gamma$ heterotrimers in both states. Our data suggest

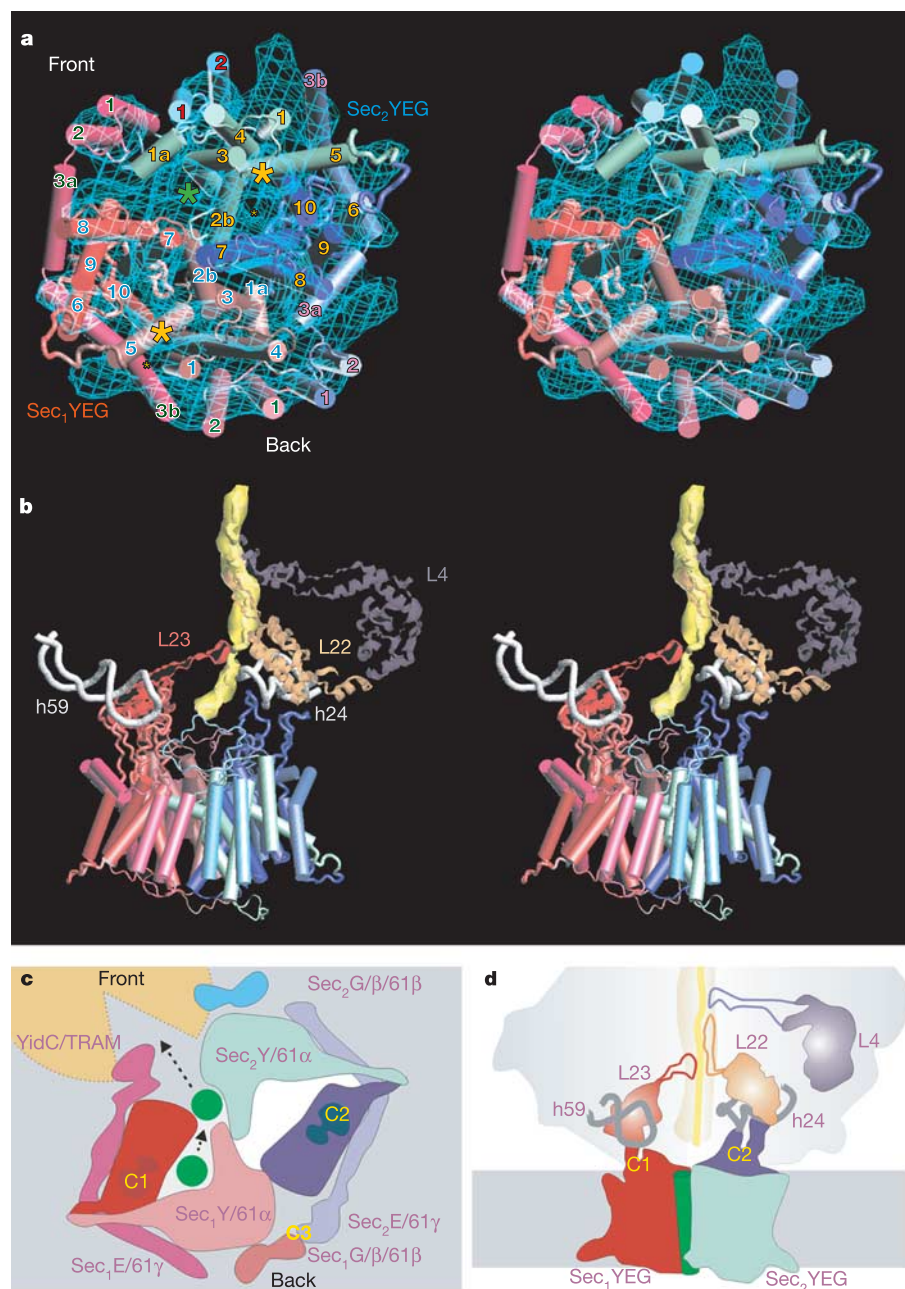


Figure 4 | The path of the nascent chain through the ribosome and PCC.

a, Stereo view of a front-to-front SecYEG_{Ec} model fitted into the translocating PCC electron microscopy density showing prominent regions of density unaccounted for (green and yellow asterisks). The PCC is viewed within the plane of the membrane, with the ribosome behind the plane, and coloured and rendered as before with the TMHs numbered.

b, Stereo view of the nascent chain (yellow rattler) fitted into the isolated polypeptide density inside the ribosome. The front view is shown. Schematic versions of the PCC at the polypeptide exit site of the ribosome in views corresponding to **a** (in panel **c**) and **b** (in panel **d**). The PCC and ribosomal elements are coloured as before, with the nascent chain (TMH signal anchor) in green and yellow. See text for discussion.

that during the translocation of a hydrophilic region of the nascent chain, neither of the two heterotrimers opens to form one central, consolidated channel, but each retains its own segregated pore. The translocating PCC density shows an asymmetry in the distribution of features, with separated rod-like features dominating at the front and contiguous lamella-like features at the back, possibly reflecting greater lipid accessibility at the front (Fig. 1c). Analysis of the atomic model of the translocating PCC supports this hypothesis. The 'hook'-shaped C-terminal SecY half in Sec₁YEG forces a path towards the front, to the bulk lipid, whereas in Sec₂YEG the 'hook' dictates a path towards the back, which is blocked to lipid by a contiguous wall formed by the TMHs of SecG and SecE (Figs 2i, j and 3a, c). A signal sequence or TMH exiting laterally from the Sec₁YEG/Sec₁61αβγ pore would interact with SecY/61α TMHs 2 and 7 (ref. 32) at the front interface between both heterotrimers, from where a TMH might then interact with YidC/TRAM (refs 34, 35) and/or diffuse into the lipid bilayer (Fig. 4a, c, d).

Our NMA results support the initial role of the plug to retain tightly juxtaposed SecY halves in the non-translocating state until the plug is displaced by the nascent chain, at which point the SecY halves can open laterally¹³ (see Methods and Supplementary Results 1.2). We suggest that the partially hydrophobic nature of the plug drives its equilibrium state towards burial between linked SecY halves when not sterically obstructed by the presence of a nascent chain, ensuring the maintenance of the permeability barrier even if linked SecY halves are not tightly juxtaposed. Thus, while the nascent chain is being looped through the pore in one heterotrimer, the pore in the other heterotrimer may remain 'plugged'. The frontal alignment of the ribosome/PCC opening with the partition site of nascent chain TMHs at the PCC/membrane interface may ensure the reproducible and efficient folding and release of nascent membrane proteins. This alignment provides a dedicated site for the association of the nascent chain and ribosome–PCC complex with cytosolic proteins, such as chaperones and SecA (refs 36–38), through the frontal opening (~20 × 40 Å; see Supplementary Discussion 2.1) and membrane proteins located at the front PCC/membrane interface. The observation of possible interactions of the projections of ribosomal proteins L4, L22 and L23 with the nascent chain is intriguing and may suggest some form of nascent-chain-dependent communication between the ribosome and the PCC.

METHODS

Generation of the RNC–SecYEG complex. *E. coli* ribosomes were programmed in a cell-free translation system with mRNA encoding a chimaeric nascent chain (Supplementary Fig. 1), consisting of an N-terminal Strep-tag followed by the first 74 residues of *E. coli* FtsQ, which includes the TMH signal anchor, and residues 132–170 of *E. coli* SecM, which includes the 17-amino-acid SecM stalling sequence¹⁸. A nascent chain consisting of FtsQ residues 1–108 has previously been crosslinked to SecY during co-translational translocation³⁶, and the SecM stalling sequence has been shown to interact tightly with the ribosomal polypeptide exit tunnel while efficiently arresting ribosome elongation¹⁸. Thus, stable RNCs were generated with nascent chains of a uniform length and sequence without using truncated mRNA. After *in vitro* translation, the RNCs were affinity-purified using the N-terminal Strep-tag and concentrated by ribosomal pelleting (Supplementary Fig. 1b–d). SecYEG was over-expressed in *E. coli*, detergent-solubilized and purified to homogeneity by standard techniques. The RNC–SecYEG complex was reconstituted and analysed (Supplementary Fig. 1e). The binding affinity for detergent-solubilized SecYEG was estimated to be tenfold higher for a ribosome containing a nascent chain compared to an empty ribosome. The presence of the nascent chain thus proved to be essential, because, in contrast to eukaryotes, empty *E. coli* ribosomes do not have sufficient affinity (apparent $K_d \sim 2 \mu\text{M}$) for the detergent-solubilized PCC to form ribosome–SecYEG complexes under the conditions used for cryo-EM (32 mM). The sample used for cryo-EM contained a tenfold molar excess of SecYEG over RNCs.

Electron microscopy, image processing and fitting of atomic ribosome models. Samples were applied to carbon-coated holey grids as published³⁹. Micrographs were recorded under low-dose conditions on a Philips/FEI Tecnai F30 field-emission gun electron microscope at 300 kV using a defocus range of

1.5–4.3 μm, and scanned on a Zeiss/Intergraphics Scanner (Zeiss/Intergraphics Imaging), corresponding to a pixel size of 3.59 Å on the object scale. The data were analysed using the SPIDER software package⁴⁰. After automated particle picking followed by visual inspection, supervised classification⁴¹ was performed on the starting set of 104,257 picked particles. Particles (53,325 in total) corresponding to a translocating PCC bound to the RNC (PCC occupancy 100%) were used for the final, CTF-corrected reconstruction. The fall-off of Fourier amplitudes towards higher spatial frequencies was corrected using the X-ray solution scattering intensity distribution of 70S ribosomes from *E. coli* (ref. 42) during each round of refinement. Refinement of an *E. coli* ribosome atomic model into the cryo-EM reconstruction was performed using the program RSRef (ref. 43), a real-space refinement module for the TNT program⁴⁴. An almost-complete atomic *E. coli* ribosome model obtained by RSRef was adapted from ref. 45, and complemented with the addition of RNA and protein models from the X-ray structure of the *Deinococcus radiodurans* large ribosomal subunit⁴⁶. **Normal mode-based flexible fitting of SecYEG models.** NMA with an elastic network⁴⁷ was performed on the X-ray structure of the SecYEG heterotrimer from *Methanococcus jannaschii* (ref. 13). Residues corresponding to the plug (SecY: 44–68, *M. jannaschii*; 40–75, *E. coli*) and the Secβ subunit were removed for observation of low-frequency, high-amplitude normal modes corresponding to the opening of SecY halves and the motions of loops (Supplementary Fig. 2). Secβ, which lies at the periphery of the SecYEG_{Mj} structure and makes few interchain contacts, contributes to high-frequency normal modes and masks the low-frequency normal modes corresponding to interdomain movement. Therefore, this chain needed to be removed. For normal mode-based flexible fitting (NMFF), a model of *E. coli* SecYEG (SecYEG_{Ec}) was generated by homology-modelling (similarity/identity between SecY of *M. jannaschii* and *E. coli* (40%/20%), and *M. jannaschii* and mouse (34%/55%)) based on SecYEG_{Mj}. The additional helices of *E. coli* were modelled according to the placement of helices determined in ref. 13 into the two-dimensional electron crystal structure of *E. coli* SecYEG determined in ref. 48. Only the ten lowest-frequency normal modes were used to direct flexible fitting of atomic models into the PCC electron microscopy density. In order to observe normal modes corresponding to the opening of SecY halves, models of plug-less SecYEG_{Ec} were used in which the two linked SecY halves in each complex were separated by 1 Å on the hingeless side. Only Cα coordinates were used for NMFF and NMA, as density for sidechains is not well resolved in the current moderate-resolution electron microscopy map. The models obtained upon NMFF were energy-minimized using CNS version 1.1 (ref. 49). The reliability of the PCC models generated by NMFF extends to the orientation of domains and the general arrangement of helices within these domains, but not to the exact positioning of individual helices and residues or the conformation of loops.

Received 17 May; accepted 8 August 2005.

- Simon, S. M. & Blobel, G. A protein-conducting channel in the endoplasmic reticulum. *Cell* **65**, 371–380 (1991).
- Wickner, W., Driessen, A. J. M. & Hartl, F. U. The enzymology of protein translocation across the *Escherichia coli* plasma membrane. *Annu. Rev. Biochem.* **60**, 101–124 (1991).
- Brundage, L. *et al.* The purified *E. coli* integral membrane protein SecY/E is sufficient for reconstitution of SecA-dependent precursor protein translocation. *Cell* **62**, 649–657 (1990).
- Gorlich, D. & Rapoport, T. A. Protein translocation into proteoliposomes reconstituted from purified components of the endoplasmic reticulum membrane. *Cell* **75**, 615–630 (1993).
- Gilmore, R. & Blobel, G. Translocation of secretory proteins across the microsomal membrane occurs through an environment accessible to aqueous perturbants. *Cell* **42**, 497–505 (1985).
- Simon, S. M., Blobel, G. & Zimmerberg, J. Large aqueous channels in membrane vesicles derived from the rough endoplasmic reticulum of canine pancreas or the plasma membrane of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **86**, 6176–6180 (1989).
- Mothes, W. *et al.* Molecular mechanism of membrane protein integration into the endoplasmic reticulum. *Cell* **89**, 523–533 (1997).
- Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433**, 377–381 (2005).
- Beckmann, R. *et al.* Alignment of conduits for the nascent polypeptide chain in the ribosome–Sec61 complex. *Science* **278**, 2123–2126 (1997).
- Menetret, J.-F. *et al.* The structure of ribosome–channel complexes engaged in protein translocation. *Mol. Cell* **6**, 1219–1232 (2000).
- Beckmann, R. *et al.* Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107**, 361–372 (2001).
- Morgan, D. G. *et al.* Structure of the mammalian ribosome–channel complex at 17 Å resolution. *J. Mol. Biol.* **324**, 871–886 (2002).
- van den Berg, B. *et al.* X-ray structure of a protein-conducting channel. *Nature* **427**, 36–44 (2004).

14. Rapoport, T. A., Goder, V., Heinrich, S. U. & Matlack, K. E. Membrane-protein integration and the role of the translocation channel. *Trends Cell Biol.* **14**, 568–575 (2004).
15. Prinz, A. *et al.* Evolutionarily conserved binding of ribosomes to the translocation channel via the large ribosomal tRNA. *EMBO J.* **19**, 1900–1906 (2000).
16. Raden, D., Song, W. & Gilmore, R. Role of the cytoplasmic segments of Sec61 α in the ribosome-binding and translocation-promoting activities of the Sec61 complex. *J. Cell Biol.* **150**, 53–64 (2000).
17. Cheng, Z., Jiang, Y., Mandon, E. C. & Gilmore, R. Identification of cytoplasmic residues of Sec61p involved in ribosome binding and cotranslational translocation. *J. Cell Biol.* **168**, 67–77 (2005).
18. Nakatogawa, H. & Ito, K. The ribosomal exit tunnel functions as a discriminating gate. *Cell* **108**, 629–636 (2002).
19. Bessonneau, P., Besson, V., Collinson, I. & Duong, F. The SecYEG preprotein translocation channel is a conformationally dynamic and dimeric structure. *EMBO J.* **21**, 995–1003 (2002).
20. van der Sluis, E. O., Nouwen, N. & Driessen, A. J. M. SecY–SecY and SecY–SecG contacts revealed by site-specific crosslinking. *FEBS Lett.* **527**, 159–165 (2002).
21. Tama, F., Miyashita, O. & Brooks, C. L. III NMFF: Flexible high-resolution annotation of low-resolution experimental data from cryo-EM maps using normal mode analysis. *J. Struct. Biol.* **147**, 315–326 (2004).
22. Go, N., Noguti, T. & Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl Acad. Sci. USA* **80**, 3696–3700 (1983).
23. Kaufmann, A. *et al.* Cysteine-directed cross-linking demonstrates that helix 3 of SecE is close to helix 2 of SecY and helix 3 of a neighbouring SecE. *Biochemistry* **38**, 9115–9125 (1999).
24. Veenendaal, A., van der Does, C. & Driessen, A. Mapping the sites of interaction between SecY and SecE by cysteine scanning mutagenesis. *J. Biol. Chem.* **276**, 32559–32566 (2001).
25. Tani, K., Tokuda, H. & Mizushima, S. Translocation of ProOmpA possessing an intramolecular disulfide bridge into membrane vesicles of *Escherichia coli*. Effect of membrane energization. *J. Biol. Chem.* **265**, 17341–17347 (1990).
26. Wirth, A. *et al.* The Sec61p complex is a dynamic precursor activated channel. *Mol. Cell* **12**, 261–268 (2003).
27. Martoglio, B., Hofmann, M. W., Brunner, J. & Dobberstein, B. The protein-conducting channel in the membrane of the endoplasmic reticulum is open laterally toward the lipid bilayer. *Cell* **81**, 207–214 (1995).
28. Levy, R., Wiedmann, M. & Kreibich, G. *In vitro* binding of ribosomes to the β subunit of the Sec61p protein translocation complex. *J. Biol. Chem.* **276**, 2340–2346 (2001).
29. Nishiyama, K.-i., Mizushima, S. & Tokuda, H. A novel membrane protein involved in protein translocation across the cytoplasmic membrane of *Escherichia coli*. *EMBO J.* **12**, 3409–3415 (1993).
30. Schatz, P. J. *et al.* One of three transmembrane stretches is sufficient for the functioning of the SecE protein, a membrane component of the *E. coli* secretion machinery. *EMBO J.* **10**, 1749–1757 (1991).
31. Kalies, K. U., Rapoport, T. A. & Hartmann, E. The β subunit of the Sec61 complex facilitates cotranslational protein transport and interacts with the signal peptidase during translocation. *J. Cell Biol.* **141**, 887–894 (1998).
32. Plath, K. *et al.* Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell* **94**, 795–807 (1998).
33. Laird, V. & High, S. Discrete cross-linking products identified during membrane protein biosynthesis. *J. Biol. Chem.* **272**, 1983–1989 (1997).
34. Scotti, P. A. *et al.* YidC, the *E. coli* homologue of mitochondrial Oxa1p, is a component of the Sec translocase. *EMBO J.* **19**, 542–549 (2000).
35. High, S. *et al.* Site-specific photocross-linking reveals that Sec61p and TRAM contact different regions of a membrane-inserted signal sequence. *J. Biol. Chem.* **268**, 26745–26751 (1993).
36. Valent, Q. A. *et al.* The *Escherichia coli* SRP and SecB targeting pathways converge at the translocon. *EMBO J.* **17**, 2504–2512 (1998).
37. Neumann-Haefelin, C., Schafer, U., Muller, M. & Koch, H. G. SRP-dependent cotranslational targeting and SecA-dependent translocation analyzed as individual steps in the export of a bacterial protein. *EMBO J.* **19**, 6419–6426 (2000).
38. Zito, C. R. & Oliver, D. Two-stage binding of SecA to the bacterial translocon regulates ribosome–translocon interaction. *J. Biol. Chem.* **278**, 40640–40646 (2003).
39. Wagenknecht, T., Grassucci, R. & Frank, J. Electron microscopy and computer image averaging of ice-embedded large ribosomal subunits from *Escherichia coli*. *J. Mol. Biol.* **199**, 137–147 (1988).
40. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
41. Valle, M. *et al.* Cryo-EM reveals an active role for aminoacyl-tRNA in the accommodation process. *EMBO J.* **21**, 3557–3567 (2002).
42. Gabashvili, I. S. *et al.* Solution structure of the *E. coli* 70S ribosome at 11.5 Å resolution. *Cell* **100**, 537–549 (2000).
43. Chapman, M. S. Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron density function. *Acta Crystallogr. A* **51**, 69–80 (1995).
44. Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Crystallogr. A* **43**, 489–501 (1987).
45. Gao, H. *et al.* Study of the structural dynamics of the *E. coli* 70S ribosome using real-space refinement. *Cell* **113**, 789–801 (2003).
46. Harms, J. M. *et al.* Alterations at the peptidyl transferase centre of the ribosome induced by the synergistic action of the streptogramins dalbapristin and quinupristin. *BMC Biol.* **2**, 4 (2004).
47. Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908 (1996).
48. Breyton, C. *et al.* Three-dimensional structure of the bacterial protein-translocation complex SecYEG. *Nature* **418**, 662–665 (2002).
49. Brunger, A. T. *et al.* Crystallography NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. A. Grassucci for training on the Tecnai F30 electron microscope; G. S. Allen for assistance with supervised classification, the software package O, and discussion of the manuscript; and M. Watters for assistance with the illustrations. N.B. thanks I. Berger and T. Ishikawa for advice on sample preparation and cryo-EM density symmetry analysis. This work was supported by Howard Hughes Medical Institute, National Science Foundation and National Institutes of Health (NIH) grants to J.F., and a Multiscale Modeling Tools for Structural Biology grant, funded by the NIH, to C.L.B. III. N.B. was supported by a grant from the Swiss National Science Foundation (SNSF), the NCCR Structural Biology program of the SNSF and a Young Investigator grant from the Human Frontier Science Program. C.S. was supported by post-doctoral fellowships from the Roche Research Foundation and the Ernst Schering Research Foundation.

Author Contributions Grid preparation, cryo-EM, data processing, and atomic model generation, fitting, refinement and interpretation were done by K.M. (in the laboratory of J.F.). *E. coli* SecYEG–RNC complex preparation and PCC electron microscopy density symmetry analysis was performed by C.S. and S.J. (in the laboratory of N.B.). T.S. (J.F.) assisted in cryo-EM data processing. F.T. (in the laboratory of C.L.B. III) performed the NMA and NMFF.

Author Information Coordinates for the translocating and non-translocating PCC have been deposited in the RCSB Protein Data Bank, with the accession codes 2AKI and 2AKH, respectively. The cryo-EM map of the *E. coli* RNC–SecYEG complex has been deposited in the EBI Macromolecular Structure Database, with the accession code EMD-1143. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.F. (Joachim@wadsworth.org).

A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome

Jeremiah R. Wagner¹, Joseph S. Brunzelle³, Katrina T. Forest² & Richard D. Vierstra¹

Phytochromes are red/far-red light photoreceptors that direct photosensory responses across the bacterial, fungal and plant kingdoms. These include photosynthetic potential and pigmentation in bacteria as well as chloroplast development and photomorphogenesis in plants. Phytochromes consist of an amino-terminal region that covalently binds a single bilin chromophore, followed by a carboxy-terminal dimerization domain that often transmits the light signal through a histidine kinase relay. Here we describe the three-dimensional structure of the chromophore-binding domain of *Deinococcus radiodurans* phytochrome assembled with its chromophore biliverdin in the Pr ground state. Our model, refined to 2.5 Å resolution, reaffirms Cys 24 as the chromophore attachment site, locates key amino acids that form a solvent-shielded bilin-binding pocket, and reveals an unusually formed deep trefoil knot that stabilizes this region. The structure provides the first three-dimensional glimpse into the photochromic behaviour of these photoreceptors and helps to explain the evolution of higher plant phytochromes from prokaryotic precursors.

Prokaryotes and eukaryotes employ a complex array of photoreceptors to coordinate their response to the ambient light environment. One of the most influential is the phytochrome superfamily, a large and diverse group of photoreceptors that use a bilin (or linear tetrapyrrole) chromophore for light detection^{1,2}. These photoreceptors bind bilins via a thioether linkage to a cysteine within the polypeptide, using an intrinsic lyase activity. After assembly, phytochromes sense red and far-red light through two relatively stable conformers, a red-light-absorbing Pr form and a far-red-light-absorbing Pfr form. By photoconverting between active and inactive forms, phytochromes act as light-regulated master switches in numerous signalling cascades. Phytochrome photoreceptors were first discovered in higher plants because of their ability to initiate red/far-red photoresponses of agricultural significance such as seed germination and flowering (reviewed in ref. 3). More recently, they have been found in cyanobacteria, proteobacteria, actinobacteria, fungi and slime moulds^{1,4}. Despite their importance, we still do not fully understand the initial molecular events that allow phytochromes to switch reversibly between the Pr and Pfr forms, nor how this transition is transduced to appropriate sensory cascades. In an attempt to define how phytochromes work at the atomic level, we have solved the three-dimensional structure of the light-sensing region of a phytochrome from the bacterium *Deinococcus radiodurans* complexed with its native chromophore, biliverdin.

Structure determination and overall fold

The first 321 residues of *D. radiodurans* bacteriophytochrome photoreceptor (DrBphP) constitute the chromophore-binding domain (DrCBD) and provide a stable, soluble and spectrally active fragment of phytochrome that is amenable to structural studies (Supplementary Fig. 1)^{4,5}. Although DrCBD readily assembles with biliverdin to generate a protein with a normal Pr spectrum⁴, its Pfr spectrum is substantially bleached, in accord with the need of a downstream PHY

domain to stabilize the Pfr conformer^{4,6,7}. The crystal structure of DrCBD was determined by two-wavelength anomalous dispersion methods, as described in Supplementary Table 1 and the Methods.

As predicted by PFAM⁸ and sequence alignments^{1,4,6}, two well-known domain folds were easily identified in the DrCBD structure (Fig. 1). The PAS (Per/Arndt/Sim) domain encompasses residues 38 to 128 with a five-stranded antiparallel β -sheet (β 2, β 1, β 5, β 4 and β 3) flanked on one side by three α helices (α 1– α 3). Its concave front surface is possibly a protein–protein signalling interface, on the basis of comparisons with other PAS domains bound to protein partners^{9,10}. Following the PAS domain is a GAF (cGMP phosphodiesterase/adenyl cyclase/FhlA) domain, confirmed biochemically to form most of the bilin-binding pocket^{1,4,6}. This domain contains a six-stranded antiparallel β sheet (β 9, β 10, β 11, β 6, β 7 and β 8) sandwiched between a three-helix bundle (α 4, α 5, α 8) and α 6 and α 7. As in most members of the phytochrome superfamily, the PAS domain of DrCBD is preceded by an $\sim$ 35-residue random-coil. Within the bacterial and fungal phytochrome clades, this extension contains the cysteine that covalently binds the A ring of the bilin chromophore^{4,11}.

Knotted interface

The PAS and GAF domains are connected by a 10-residue linker (residues 129–138) and limited electrostatic and hydrophobic interactions at the domain interface (for example, Asp 95 to Arg 218; Arg 100 to Asp 300 and Thr 303; Gln 53 to Thr 246; His 219C ϵ 1 to Ala 96C β). The most extensive interface between the domains, however, is also the most surprising feature of the DrCBD structure. A deep trefoil knot is formed as the 35 N-terminal residues upstream of the PAS domain (plus 13 residues of the T7 tag) pass through the loop between β 9 and α 7 (Fig. 1a, c). This loop (residues 225–257) is an insertion within the otherwise structurally conserved GAF domain. The main-chain hydrogen-bonding pattern of the knot creates a small antiparallel β -sheet-like structure in which the outer

¹Department of Genetics, ²Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin 53706 USA. ³Life Sciences Collaborative Access Team, Northwestern University, Argonne, Illinois 60439, USA.

two mini-strands are donated by the GAF domain loop, with residues from the N-terminal region of the CBD forming the central mini-strand (Figs 1c and 2a). Partly owing to a *cis*-peptide bond at residue 235, this $\beta 2'$ secondary structure forms an unusual continuation of the β -sheet from the PAS domain $\beta 2$. The interface extends to the GAF domain β -sheet by using the chromophore as a stepping stone (Arg 254 from $\beta 3'$ to the propionate side chain of biliverdin ring B and from the propionate side chain of biliverdin ring C to Ser 272 and Ser 274 of $\beta 10$) (Fig. 2a). The knot is stabilized by a small but critical hydrophobic core (Fig. 2b). The key to this knot is Ile 35; it is

positioned inside the loop and makes van der Waals contacts with side chains of Leu 41 and loop residues Val 232, Leu 234, Leu 248 and Leu 253, potentially explaining why Ile 35 is necessary for protein folding in pea Phy¹² and *Agrobacterium tumefaciens* BphP2 (ref. 4). The side chain of Gln 36 from the N-terminal polypeptide spans the base of the GAF-inserted lasso and makes hydrogen bonds to main-chain atoms of Ala 225 and Arg 254, just at the entrance and exit from the GAF domain into the lasso (Fig. 2b).

Although it was once considered impossible, several proteins have been identified recently with overhand, or trefoil, knots in the polypeptide chain^{13–17}. The DrCBD contains a rare deep trefoil knot in which 48 amino acids protrude from the crossover. Given the long-held dogma that proteins do not fold into knots, we examined experimental electron density maps (Fig. 2c) and omit maps to determine whether alternative connectivities were possible; none were found. Neither do inadvertent chain tracing through symmetry mates nor domain-swapping scenarios explain the unexpected topology¹⁸. Φ/Ψ angles, temperature factors and the real space correlation coefficient¹⁹ throughout this region are representative of values for the entire structure. Collectively, these data give us confidence that the knot is authentic.

Bilin chromophore structure and ligation

Biliverdin within DrCBD assumes a partially extended conformation with the best fit to the electron density supporting a $5Z_{\text{syn}}, 10Z_{\text{syn}}, 15Z_{\text{anti}}$ configuration for the methylene linkers connecting the four pyrrole rings (Fig. 3a, b). The Z_{syn} configuration between the B and C rings and the Z_{anti} configuration between the C and D rings are in agreement with several resonance Raman spectroscopy studies of the phytochromobilin (PFB) chromophore within plant phytochromes^{20–22} as well as crystal structures of the phycocyanobilin (PCB) chromophore within the photosynthetic light-harvesting complex^{23,24}. However, for the A–B linkage, the $5Z_{\text{syn}}$ conformation in our structure differs from the extended $5Z_{\text{anti}}$ conformation suggested for those systems. As is the case for bilins in solution, the A, B and C rings of biliverdin are coplanar in DrCBD. The D ring is rotated 44° away from this plane (Fig. 1b). This rotation, which may strain the C15 = C16 double bond, appears to be stabilized by a hydrogen bond from the D-ring carbonyl oxygen to His 290 and by hydrophobic contacts of the D-ring vinyl group with Phe 198 and Phe 203 (Fig. 3c).

Bacterial phytochromes such as DrBphP bind biliverdin via a thioether linkage between a conserved cysteine upstream of the PAS domain and the A-ring C3 vinyl group^{4,11}. Our structure confirms this linkage and suggests a covalent bond to the terminal C3² carbon of the vinyl group, thus forming an extended linker between the polypeptide backbone and the bilin (Fig. 3). The intrinsic bilin lyase mechanism of phytochromes has not been resolved, although the protein structure surrounding the Cys24 binding site suggests a possible mechanism. The side-chain carboxylates of Glu 25 and Glu 27 are 4.9 and 7.2 Å, respectively, from the Cys24 sulphur. In the apoprotein form of this Phy, either or both residues could approach and activate the thiol for subsequent nucleophilic attack at C3². The geometry of the site and the potential for electron resonance within biliverdin may favour this reaction with the C3² carbon over the C3¹ carbon, which is the attachment site for the more reduced bilins in plant and cyanobacterial phytochromes (Fig. 3a). The final resolution of the attachment mechanism for biliverdin will require additional biochemical experiments.

Chromophore-binding pocket

The unique photochromicity of phytochromes results from key interactions between the bilin and its binding pocket^{1,6,22}. Biliverdin slides into a crevice formed between $\alpha 6$ and $\alpha 7$ on one side and the GAF β -sheet on the other, largely burying the extended bilin within the protein (Figs 1b, 3c and 4a). The PAS domain and upstream

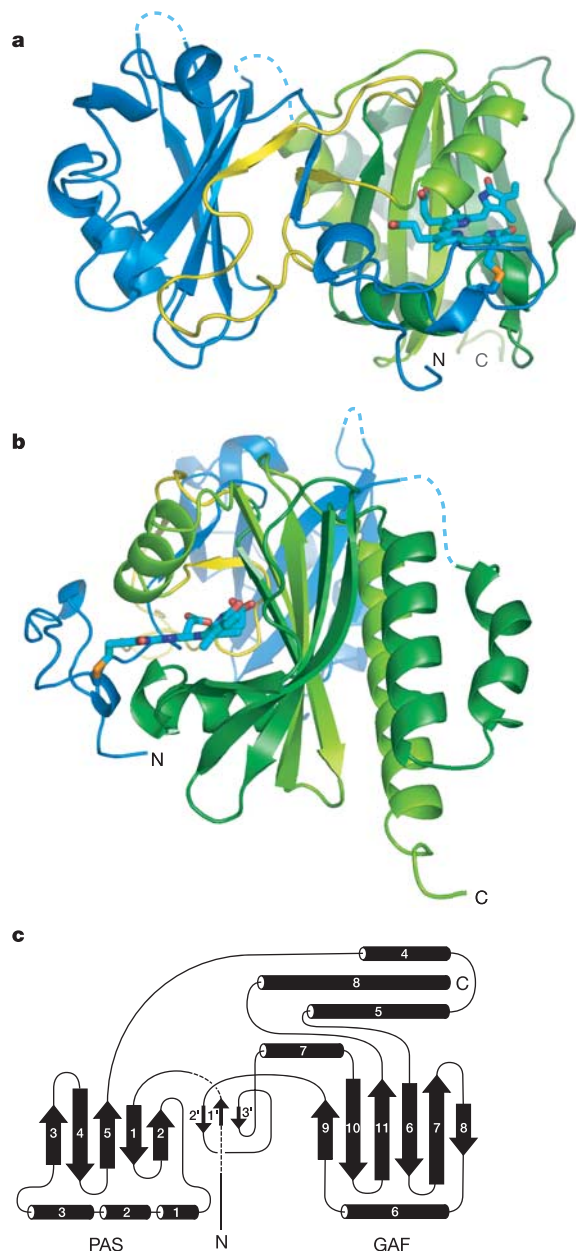


Figure 1 | Three-dimensional structure of DrCBD. **a**, The DrCBD fold consists of the N-terminal region and PAS domain (sky blue) and the GAF domain (light green for first half of sheet, yellow for lasso and dark green for second half of sheet). Biliverdin (blue) is covalently attached via thioether linkage to Cys24 (orange). Dashed lines identify disordered residues missing from the model. **b**, The nonplanar biliverdin rests in a binding groove within the GAF domain. Colours as in **a**, with model rotated $\sim 90^\circ$ about the y axis. **c**, Topology of CBD, including the knotted mini-sheet (centre). Numbers refer to secondary structure elements.

sequence provide little direct contact except for residues immediately surrounding the ligating Cys 24.

Consistent with previous binding studies with bilin analogues in plant phytochromes²⁵, the propionic-acid side chains of the B and C rings are particularly important for positioning biliverdin. These side chains penetrate deeply within the cleft between $\alpha 7$ and β -strands 9 and 10, and are fixed by highly conserved charge interactions, direct and water-mediated hydrogen bonds, and hydrophobic interactions (Fig. 3c). In particular, the carbonyl oxygens of the B-ring propionate form salt bridges with the amines of Arg 254 and form hydrogen bonds with the hydroxyl group of Tyr 216 and main-chain nitrogen of Ser 257, whereas the C-ring propionate oxygens hydrogen-bond to His 260 and the hydroxyl groups of Ser 272 and Ser 274. A water molecule (Wat 18) forms a water-mediated bridge between the C-ring propionate and His 290. There are also van der Waals interactions between Ile 29 and the aliphatic carbons of the B-ring propionate.

Other interactions contribute to biliverdin orientation in the pocket. Asp 207, Ile 208 and Pro 209 (the DIP motif) create a kink in the polypeptide chain between $\beta 8$ and $\alpha 6$, allowing these residues access to the chromophore (Fig. 3c). The side chain of Asp 207 forms a hydrogen bond with its own main-chain nitrogen to stabilize this proline-induced kink, and forms a bifurcated hydrogen bond to the nitrogens of the A and B rings. There are hydrophobic interactions between Ile 208 and the C ring, between Pro 209 and the C10 carbon that links the B and C rings and between Pro 209 and the B ring. His 260, which was previously shown to be critical for biliverdin conjugation to DrBphP^{4,5}, forms hydrogen bonds with the pyrrole nitrogens of rings B and C and its planar shape helps to fill a van der

Waals sandwich between these planar rings and Leu 264. Wat 12 mediates a remarkable hydrogen-bonding network from His 260N δ 1 to the pyrrole nitrogens of biliverdin rings A, B and C (Fig. 3c).

During phototransformation from Pr to Pfr, a substantial re-orientation of ring D is expected^{22,26}. A notable feature in our model of DrCBD is the relatively sparse packing of side chains around the D ring to allow this rotation. Hydrophobic residues Met 174, Val 186 and Leu 286 line the pocket but are not tightly packed against the D ring. The main contacts to the D ring are polar on one edge (a hydrogen bond from the ring carbonyl to His 290N ϵ 2) and hydrophobic on the other (Tyr 176, Tyr 263, Phe 198 and Phe 203) (Fig. 3b). Although Tyr 176 is buried under the chromophore in our Pr structure, the latter three aromatic residues are partially solvent-exposed, and the phenylalanine side chains contact the D-ring vinyl group.

Insights into the Pr to Pfr photochemistry

The crystal structure of DrCBD provides insight into the unique red/far-red photochromicity of phytochromes. There is general agreement, based on spectroscopic results, that phototransformation to Pfr in plant and bacterial phytochromes involves a $15Z_{anti}$ to $15E_{anti}$ isomerization of the C15 = C16 double bond (Fig. 3a)^{20,22,27,28}. Space around the D ring would allow rotation relative to the other pyrrole rings during the Pr to Pfr phototransformation with minimal interference from amino acids lining the pocket and only the clash of eclipsing methyl carbons from the C and D rings of the chromophore limiting rotation angle in either direction (Fig. 3c). Pre-existing torsional strain on the Pr D ring may serve to reduce the activation energy for rotation. The C15 = C16 isomerization would dissolve the

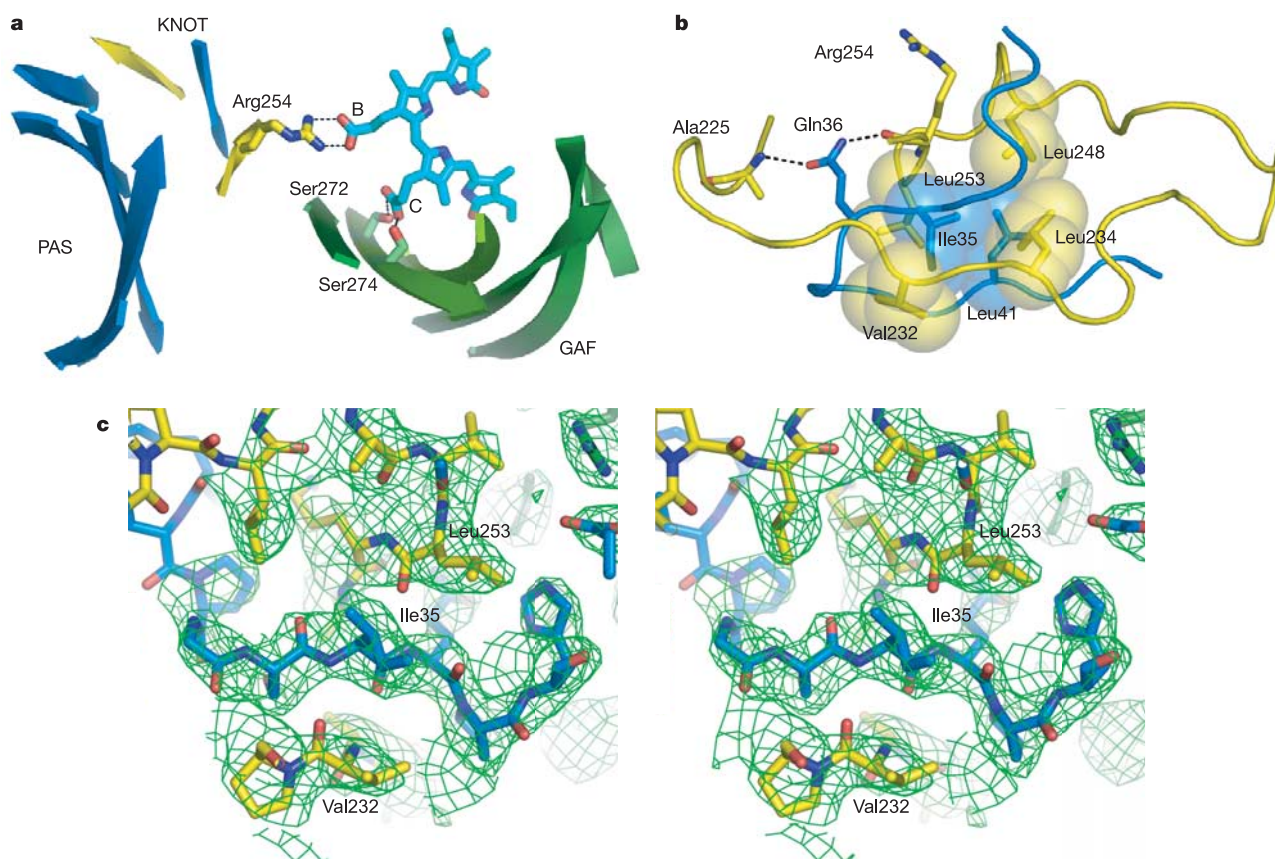


Figure 2 | Structure of the deep trefoil knot. **a**, A semi-continuous β -sheet 'snake' is formed by the PAS (blue) and GAF (green) domains and the polypeptide knot, with Arg 254, Ser 272 and Ser 274 stepping across the chromophore via propionate side chains of rings B and C. **b**, Conserved side

chains from both the N-terminal extension (blue) and the GAF-inserted lasso (yellow) form the small hydrophobic core within the knot. **c**, Stereo depiction of the 'knot' region of the final refined model superimposed on the experimental F_o electron density map, contoured in green at 1.0σ .

interaction of the D ring with its Pr bonding partner His 290 and might perturb Phe 198 and Phe 203 by removing favourable van der Waals interactions.

In addition to rapid isomerization of the C15 = C16 bond, a subset of published experiments have predicted a torsion angle rotation from the 15E_{anti} to the 15E_{syn} configuration at the C14–C15 bond^{20,22,27,28}. However, recent elegant studies with chemically locked biliverdin analogues demonstrated that the locked 15E_{anti} conformer of biliverdin covalently associates with a bacterial phytochrome and exhibits spectral properties similar to the Pfr form²⁸. If this further 15E_{anti} to 15E_{syn} movement does occur, it would cause a steric clash between the D ring and Tyr 176 (if the rotation is away from the solvent) or Tyr 263 (if the rotation is towards the solvent). A continued rotation to the 15E_{syn} position could also lead to a clash with the side chain of Asp 207 (Fig. 3c).

The rotation(s) of biliverdin lead to a partial decoupling of the π conjugation system and a red-shift in the absorption spectrum of phytochrome that is characteristic of the Pfr conformer²⁹. Presumably, such movement(s) induce additional conformational changes in the polypeptide to generate the distinctive shape of the Pfr conformer, which for many fungal and prokaryotic phytochromes affects the activity of an appended kinase domain^{1,4}. For plant phytochromes Pfr formation exposes a nuclear localization signal in a downstream PAS-related domain³⁰ or uncovers interaction motifs for downstream effectors³¹.

It is also possible that C15 = C16 bond isomerization could lead to perturbation of the solvent-exposed Phe side chains and that further torsion about C14–C15 could push Asp 207 and initiate local melting

of the DIP motif on the surface of the CBD. Either of these would lead to a change in the shape and properties of the invariant surface patch at the chromophore-binding site (Fig. 4a). A final intriguing possibility is that even slight shifts of the orientation of B and C rings within the pocket as a consequence of D-ring movement and postulated pocket structure changes could release the propionate side chains from their electrostatic interactions. This release could break the structural bridge connecting the GAF and PAS domains and thereby free the PAS domain for signal transmission (Fig. 2a). Clearly, determining the structure of the Pfr form of CBD and of *DrBphP* with an intact PHY domain downstream of the CBD will be essential to further understanding phototransformation and signal transduction by phytochromes, given the importance of the PHY domain to the establishment and maintenance of Pfr^{4,6,7}.

Functional and evolutionary implications

An important feature of the *DrCBD* structure with respect to the rest of the phytochrome superfamily is the high degree of conservation^{4,32} of essential residues that form the PAS and GAF domains, bind the bilin, and create the trefoil knot (Fig. 4a, Supplementary Fig. 2). Specific examples include the Asp-Ile-Pro (DIP) motif at the bilin-binding pocket, polar residues that interact with biliverdin propionic acid groups (Arg 254, Ser 272 and Ser 274), bulky hydrophobic residues that line the D-ring cavity (Phe 198, Phe 203, Tyr 176 and Tyr 263), and residues that form the hydrophobic core of the trefoil knot (Ile 35, Leu 41, Leu 241 and Leu 253). For example, recent work has highlighted the importance of Tyr 176 in the photochemistry of *Synechocystis* Cph1 by showing that a His substitution at this position

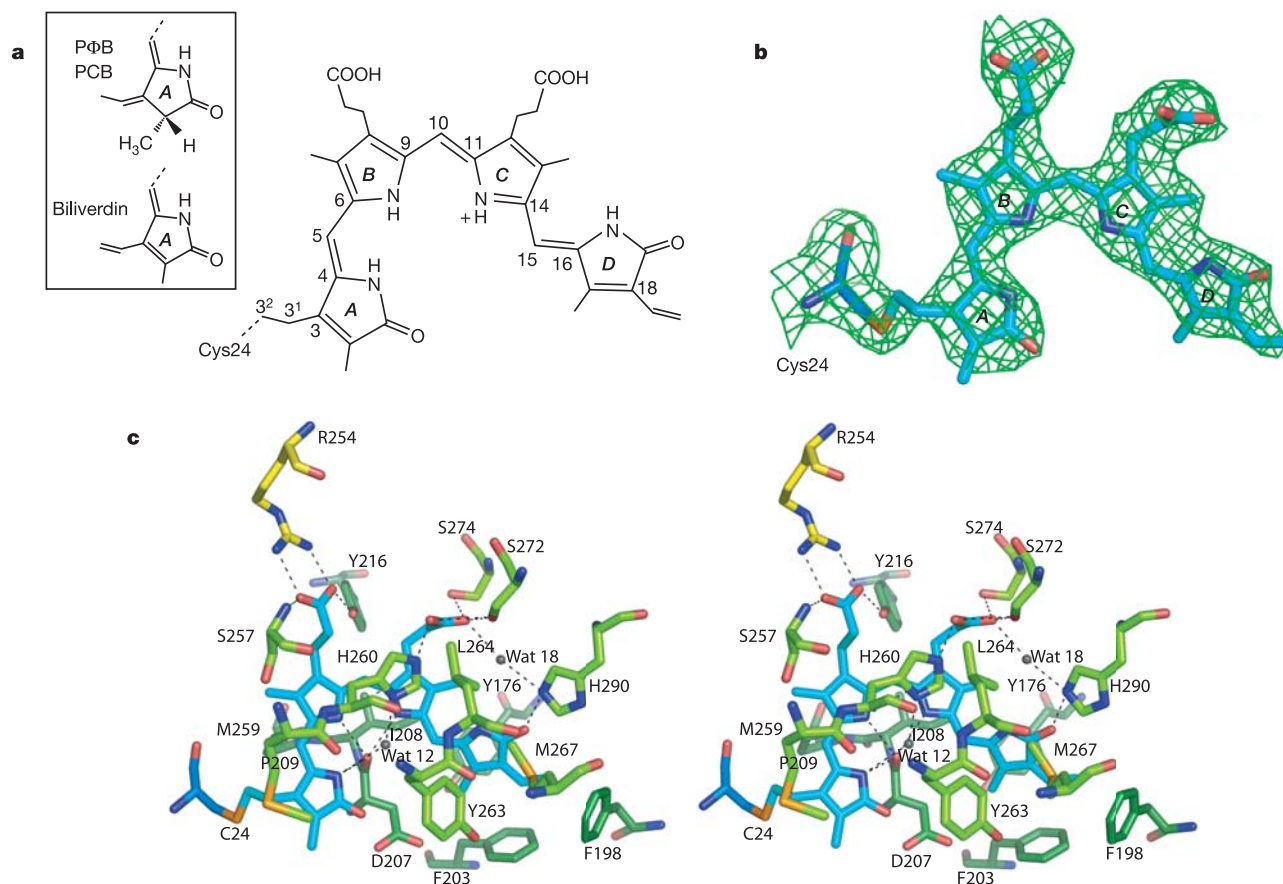


Figure 3 | Structure and linkage of biliverdin within *DrCBD*. **a**, Biliverdin, attached at C3² in the A ring, adopts a 5Z_{syn}, 10Z_{syn}, 15Z_{anti} configuration. Inset, the A-ring structure for the bilin chromophore before covalent attachment to phytochrome differs for PΦB (plants) and PCB (cyanobacteria) as compared to biliverdin (bacteria). **b**, Final refined model

of biliverdin is superimposed on the $2F_o - F_c$ electron density map contoured at 1.0σ . **c**, The biliverdin-binding pocket, shown here in stereo, is formed largely by invariant GAF domain residues (stick figure colour scheme as for Fig. 1a).

destabilizes Pfr and creates a red fluorophore³³. Consequently, we expect that our structure of this bacterial phytochrome will be highly relevant to most, if not all, members of the phytochrome superfamily.

Conservation of residues surrounding the knot, including the invariant presence of the lasso loop within the GAF domain, indicates that this unique topology is common to all red/far-red photochromic phytochromes. In fact, only the loosely defined collection of bacterial phytochrome-like proteins may lack this fold, an idea based on sequence alignments that show that this group lacks both the PAS domain and the GAF loop⁴. Although several members of this phytochrome-like group bind bilins *in vitro* (RcaE, Cika and TaxD1/PixJ)^{34–36} the resulting holoproteins are not red/far-red photoreversible, suggesting that the knot is required for this spectral characteristic. Within the holoprotein, the knot may have several functions. One may be to stabilize contacts between the PAS and GAF domains. Another may be to limit the flexibility of a photoreceptor as a mechanism to prevent undesirable energy losses due to vibration or large domain movement upon photoisomerization of biliverdin. Finally, by severely restricting movement of the N-terminal domain, the knot may orient Cys 24 for efficient conjugation to biliverdin.

How do phytochromes fold into this convoluted tertiary structure? We propose a simple co-translational mechanism in which the PAS domain folds first, leaving the N-terminal 35 amino acids

unstructured and Ile 35 solvent-exposed. As translation continues, the first half of the GAF domain, comprising the $\alpha 4$ – $\alpha 5$ helical hairpin, the $\beta 6$, $\beta 7$, $\beta 8$ half-sheet and $\alpha 6$, is translated and adopts its secondary structure. $\beta 9$ is synthesized, followed by the lasso loop. The exposed hydrophobic side chains within the lasso collapse around Ile 35 to form the compact knot core. Finally, the second half of the GAF domain is translated. Although the internal symmetry of the GAF domain is interrupted by the lasso (Fig. 1a, c), the first and last β -strands cannot zip together in the centre of the sheet until translation is complete.

With respect to the evolution of this photoreceptor class, our model may explain in molecular terms how phytochromes changed from proteobacterial photoreceptors using biliverdin as the chromophore to those present today in cyanobacteria and plants that use the more reduced bilins, PCB and P Φ B, respectively (Fig. 3a)^{1,26}. Presumably, these blue-shifted bilins offer an adaptive advantage for sensing photosynthetic competition by closely matching the absorption spectrum of chlorophyll^{3,4}. This switch required evolution of mechanisms to synthesize reduced bilins and to discriminate them from biliverdin in the lyase active site (Fig. 3a). Cyanobacterial and plant phytochromes use a cysteine in the GAF domain to bind the A-ring C3 ethylidene side chain of PCB/P Φ B^{37,38}, as opposed to the proteobacterial and (probably) fungal phytochromes that use a cysteine upstream of the PAS domain to bind the C3 vinyl side chain of biliverdin^{4,11,32}. Remarkably, modelling of this GAF cysteine into its comparable position (Met259) in *DrCBD* reveals that it would extend towards the C3 side chain, as does Cys24 in *DrCBD*, but from the opposite side of the chromophore (Fig. 4b). With no adjustment to the surrounding structure, the sulphur atom of a cysteine replacement at position 259 would be 5.8 Å from the Cys 24 sulphur and only 3.2 Å from the target C3¹ carbon of the bilin. Consequently, evolutionary transformation of the bilin attachment site could occur without structural changes, simply by acquiring a cysteine at the appropriate position in the GAF domain. In support, ref. 32 recently showed that an *Agrobacterium tumefaciens* BphP1 (Agp1) variant in which the distal cysteine was added (Val249 → Cys) preferentially assembles with PCB versus biliverdin to generate a photochromic phytochrome with typical red/far-red absorption spectra.

The structure of the bilin-binding domain of DrBphP yields the first three-dimensional insights into the photochemistry and evolution of the phytochrome superfamily. This structure provides molecular explanations for the known physico-chemical features of phytochromes, for Pr to Pfr photoconversion events that initiate responses of behavioural and agricultural significance, and for the evolution of the phytochrome family.

METHODS

Protein production and crystallization. The coding region for the first 321 (of 755 total) amino acids of DrBphP (ref. 5) was amplified from a plasmid bearing the full-length gene^{5,39} and introduced into pET21a(+) (Novagen) for expression with a 14-amino-acid N-terminal T7 tag and a C-terminal hexahistidine tag. A point mutation leading to a Pro 240 → Thr 240 substitution was inadvertently introduced during cloning, but the CBD bearing this mutation was photochemically identical to the CBD with the native sequence⁴ (Supplementary Fig. 1). Purification, crystallization and cryo-protection were performed under green safe lights. Expression, chromophore ligation in crude lysate, and purification by Ni²⁺-chelate affinity chromatography were carried out as described⁴ with the addition of an anion-exchange (Mono-Q, Amersham) step, using a 100–500 mM NaCl elution gradient. To obtain selenomethionine-derivatized protein, DrCBD apoprotein was expressed in *Escherichia coli* strain B834:pLysS (Stratagene) grown in selenomethionine expression medium (Athena Enzyme Systems/Molecular Dimensions) supplemented with 0.25 mg per litre of L-selenomethionine (Sigma)⁴⁰. Notably, even with several methionines within the bilin-binding pocket, the binding efficiency for ligand and the action spectrum of the selenomethionine-labelled protein are indistinguishable from the native protein (Supplementary Fig. 1).

Initial crystallization conditions were identified in a 1,536-condition screen

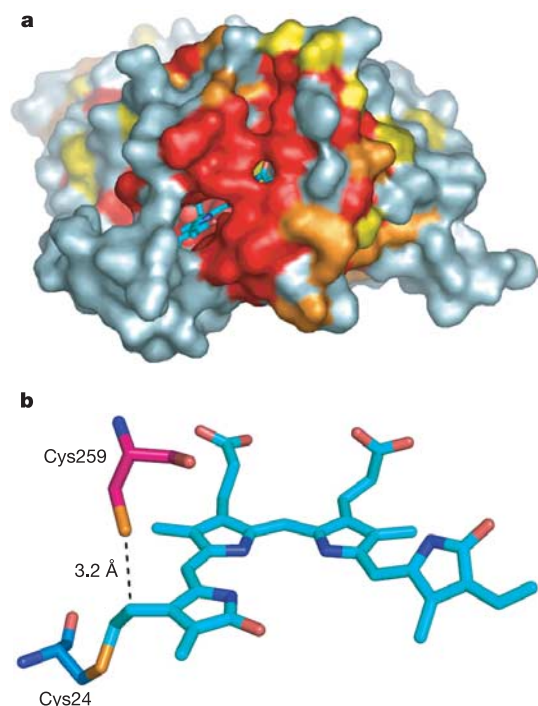


Figure 4 | Structural conservation and evolution of phytochrome photoreceptors. **a**, phytochrome superfamily conservation is apparent in this solvent-accessible surface of *DrCBD*. It is colour-coded, based on 90% (red), 75% (orange) and 60% (yellow) sequence identity among phytochromes from 17 proteobacteria, 3 cyanobacteria, 3 fungi and 17 plants. (A subset of these sequences are aligned in Supplementary Fig. 2.) Invariant residues Asp 207, Tyr 263 and Phe 203 form the bridge over the deeply buried biliverdin and the two conformations observed (although not refined) in our electron density map for Tyr 263 may play a part in allowing access to the pocket. **b**, Evolution of the plant phytochrome ligand attachment site would have required the change of a single amino acid, as shown in this view of the binding site in which the *DrCBD* Cys 24 and biliverdin (cyan) are shown with Met 259 modelled as a cysteine (Cys 259, magenta).

using microbatch under oil⁴¹ conducted by the Hauptman Woodward Institute (<http://www.hwi.buffalo.edu>). These conditions were optimized, and a recrystallization step was added to obtain diffraction-quality crystals. Briefly, large-scale sitting drops (75 µl of 30 mg ml⁻¹ protein in 30 mM pH 8.0 Tris-HCl equilibrated with an equal volume of mother liquor, consisting of 67 mM pH 4.95 sodium acetate, 3.3% PEG 400, 1 mM DTT) yielded showers of microcrystals within 24 h at room temperature. These were harvested by centrifugation, redissolved in 30 mM pH 8.0 Tris-HCl, clarified by centrifugation, and concentrated to 20 mg ml⁻¹. Macrocrystals were grown by hanging-drop vapour diffusion using 2 µl each of protein and mother liquor; spectroscopy confirmed that crystalline DrCBD is largely in the Pr state (Supplementary Fig. 1). Cryoprotection of the ~0.3 mm × 0.3 mm × 0.05 mm bright green plates in mother liquor plus 30% 2-methyl-2,4-pentanediol yielded diffraction to a maximum resolution of 2.5 Å. Native and selenomethionine-containing crystals are of the same orthorhombic space group $P2_12_12_1$ ($a = 64.9$ Å, $b = 133.7$ Å, $c = 49.9$ Å; $\alpha = \beta = \gamma = 90^\circ$) with a single molecule per asymmetric unit.

Structure determination. MAD (multiwavelength anomalous dispersion) data were collected at 98 K on a MarCCD (charge-coupled device) detector at Sector 32 ID-B at the Advanced Photon Source. Inflection and peak data sets were collected on separate crystals, and integrated and merged with HKL2000⁴² (Supplementary Table 1). The program Hyss⁴³ was used to find seven of nine possible selenium sites, five of which were refined in SHARP⁴⁴. Initial phases, with an overall figure of merit of 0.41, were improved by solvent flattening and phase extension, and provided an interpretable electron density map. Automatic model building with ARP/wARP⁴⁵ and RESOLVE⁴⁶ placed ~40 correct amino acids and a polyaniline backbone for ~50% of the protein. The remaining model was traced manually with XFIT⁴⁷ alternated with TLS and positional refinement against the 2.5 Å inflection data set using REFMAC5⁴⁸. Early in the refinement, electron density for biliverdin was apparent and the chromophore was fitted manually to the density. A modified amino acid in which Cys 24 was bonded to C3² of the A-ring vinyl group of biliverdin via a thioether linkage was refined without an energy penalty for rotation about the C15 = C16 double bond. Negative density around this bond in an electron density map calculated with Fourier coefficients F_{obs} (final 30° of data) – F_{obs} (first 30° of data) suggests that the bond is subject to radiation-induced breakage.

The final refined model includes 309 amino acid residues (residues 5–321 of DrCBD and four C-terminal His residues, with two disordered residues omitted at the tip of the β3–β4 hairpin and six in the interdomain linker), biliverdin, and 34 well-ordered water molecules. Histidine residues in the purification tag form crystal-packing interactions (His 322 to the backbone of Val 318, His 325 to symmetry-related Glu 104, and His 323 to symmetry-related Glu 130). Only two non-glycine residues, Cys 24 and His 110, fall outside the favourable or most-favourable regions of the Ramachandran plot. Ligand attachment at Cys 24 may cause backbone strain, and His 110 precedes the β-hairpin break. The aromatic ring of Phe 8 was modelled in two orientations, each having occupancy of 0.5. Other sidechains showed evidence of dual conformers but were not split owing to the limited resolution of our data. The final model has very good stereochemistry, an R_{work} of 23.7%, and an R_{free} of 26.6% for all data to 2.5 Å (Supplementary Table 1).

Structure figures were generated using Pymol (Delano Scientific, San Carlos, <http://pymol.sourceforge.net/>).

Received 27 May; accepted 9 August 2005.

- Vierstra, R. D. & Karniol, B. in *Handbook of Photosensory Receptors* (eds Briggs, W. R. & Spudich, J. L.) 171–196 (Wiley, Weinheim, 2005).
- Quail, P. H. Phytochrome phototransduction signalling networks. *Nature Rev. Mol. Cell Biol.* **3**, 85–93 (2002).
- Smith, H. Physiological and ecological functions with the phytochrome family. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **46**, 269–315 (1995).
- Karniol, B., Wagner, J. R., Walker, J. M. & Vierstra, R. D. Phylogenetic analysis of the phytochrome superfamily reveals distinct microbial subfamilies of photoreceptors. *Biochem. J.* **392**, 103–116 (2005).
- Davis, S. J., Vener, A. V. & Vierstra, R. D. Bacteriophytochromes: phytochrome-like photoreceptors from nonphotosynthetic eubacteria. *Science* **286**, 2517–2520 (1999).
- Wu, S. H. & Lagarias, J. C. Defining the bilin lyase domain: lessons from the extended phytochrome superfamily. *Biochemistry* **39**, 13487–13495 (2000).
- Cherry, J. R. et al. Carboxy-terminal deletion analysis of oat phytochrome A reveals the presence of separate domains required for structure and biological activity. *Plant Cell* **5**, 565–575 (1993).
- Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **32D**, 138–141 (2004).
- Yildiz, O. et al. Crystal structure and interactions of the Pas repeat region of

the *Drosophila* clock protein Period. *Mol. Cell* **17**, 69–82 (2005).

- Razeto, A. et al. Structure of the NcoA-1/Src-1 Pas-B domain bound to the Lxll motif of the Stat6 transactivation domain. *J. Mol. Biol.* **336**, 319–329 (2004).
- Lamparter, T. et al. The biliverdin chromophore binds covalently to a conserved cysteine residue in the N-terminus of *Agrobacterium* phytochrome Agp1. *Biochemistry* **43**, 3659–3669 (2004).
- Bhoo, S. H. et al. Phytochrome photochromism probed by site-directed mutations and chromophore esterification. *J. Am. Chem. Soc.* **117**, 11717–11718 (1997).
- Taylor, W. R. A deeply knotted protein structure and how it might fold. *Nature* **406**, 916–919 (2000).
- Nureki, O. et al. An enzyme with a deep trefoil knot for the active-site architecture. *Acta Crystallogr. D* **58**, 1129–1137 (2002).
- Taylor, W. R. & Lin, K. Protein knots: a tangled problem. *Nature* **421**, 25 (2003).
- Zarembinski, T. I. et al. Deep trefoil knot implicated in RNA binding found in an archaeobacterial protein. *Proteins Struct. Funct. Genet.* **50**, 177–183 (2003).
- Mallam, A. L. & Jackson, S. E. Folding studies on a knotted protein. *J. Mol. Biol.* **346**, 1409–1421 (2005).
- Liu, Y. & Eisenberg, D. 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**, 1285–1299 (2002).
- Vaguine, A. A., Richelle, J. & Wodak, S. J. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr. D* **55**, 191–205 (1999).
- Kneip, C. et al. Protonation state and structural changes of the tetrapyrrole chromophore during the Pr → Pfr phototransformation of phytochrome: a resonance Raman spectroscopic study. *Biochemistry* **38**, 15185–15192 (1999).
- Mroginiski, M. A. et al. Determination of the chromophore structures in the photoinduced reaction cycle of phytochrome. *J. Am. Chem. Soc.* **126**, 16734–16735 (2004).
- Braslavsky, S. E. in *Photochromisms, Molecules and Systems* (eds BrouasLaurent, H. & BrouasLaurent, D. H.) 738–755 (Elsevier Science, Amsterdam, 2003).
- Duerring, M., Schmidt, G. B. & Huber, R. Isolation, crystallization, crystal structure analysis and refinement of constitutive C-phycoerythrin from the chromatically adapting cyanobacterium *Fremyella diplosiphon* at 1.66 Å resolution. *J. Mol. Biol.* **217**, 577–592 (1991).
- Brejci, K., Ficner, R., Huber, R. & Steinbacher, S. Isolation, crystallization, crystal structure analysis and refinement of allophycocyanin from the cyanobacterium *Spirulina platensis* at 2.3 Å resolution. *J. Mol. Biol.* **249**, 424–440 (1995).
- Hanzawa, H. et al. In vitro assembly of phytochrome B apoprotein with synthetic analogs of the phytochrome chromophore. *Proc. Natl Acad. Sci. USA* **98**, 3612–3617 (2001).
- Tu, S.-L. & Lagarias, J. C. in *Handbook of Photosensory Receptors* (eds Briggs, W. R. & Spudich, J. L.) 121–149 (Wiley, Weinheim, 2005).
- Fodor, S. P., Lagarias, J. C. & Mathies, R. A. Resonance Raman analysis of the Pr and Pfr forms of phytochrome. *Biochemistry* **29**, 11141–11146 (1990).
- Inomata, K. et al. Sterically locked synthetic bilin derivatives and phytochrome Agp1 from *Agrobacterium tumefaciens* form photosensitive Pr- and Pfr-like adducts. *J. Biol. Chem.* **280**, 24491–24497 (2005).
- Gartner, W. & Braslavsky, S. E. in *Photoreceptors in Light Signaling* (ed. Baschauer, A.) 136–180 (Royal Soc. Chemistry, Cambridge, UK, 2004).
- Cheng, M., Tao, Y., Lim, J., Shaw, A. & Chory, J. Regulation of phytochrome B nuclear localization through light-dependent unmasking of nuclear-localization signals. *Curr. Biol.* **15**, 637–642 (2005).
- Ryu, J. S. et al. Phytochrome-specific type 5 phosphatase controls light signal flux by enhancing phytochrome stability and affinity for a signal transducer. *Cell* **120**, 395–406 (2005).
- Lamparter, T. et al. Biliverdin binds covalently to *Agrobacterium* phytochrome Agp1 via its ring A vinyl side chain. *J. Biol. Chem.* **278**, 33786–33792 (2003).
- Fischer, A. J. & Lagarias, J. C. Harnessing phytochrome's glowing potential. *Proc. Natl Acad. Sci. USA* **101**, 17334–17339 (2004).
- Mutsaers, M., Michel, K. P., Zhang, X. F., Montgomery, B. L. & Golden, S. S. Biochemical properties of CikA, an unusual phytochrome-like histidine protein kinase that resets the circadian clock in *Synechococcus elongatus* PCC 7942. *J. Biol. Chem.* **278**, 19102–19110 (2003).
- Terauchi, K., Montgomery, B. L., Grossman, A. R., Lagarias, J. C. & Kehoe, D. M. RcaE is a complementary chromatic adaptation photoreceptor required for green and red light responsiveness. *Mol. Microbiol.* **51**, 567–577 (2004).
- Yoshihara, S., Katayama, M., Geng, X. X. & Ikeuchi, M. Cyanobacterial phytochrome-like PixJ1 holoprotein shows novel reversible photoconversion between blue- and green-absorbing forms. *Plant Cell Physiol.* **45**, 1729–1737 (2004).
- Lagarias, J. C. & Rapoport, H. Chromopeptides from phytochrome. The structure and linkage of the Pr form of the phytochrome chromophore. *J. Am. Chem. Soc.* **102**, 4821–4828 (1980).
- Yeh, K. C., Wu, S. H., Murphy, J. T. & Lagarias, J. C. A cyanobacterial phytochrome two-component light sensory system. *Science* **277**, 1505–1508 (1997).

39. Bhoo, S. H., Davis, S. J., Walker, J., Karniol, B. & Vierstra, R. D. Bacteriophytochromes are photochromic histidine kinases using a biliverdin chromophore. *Nature* **414**, 776–779 (2001).
40. Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362**, 219–223 (1993).
41. Chayen, N. E., Shaw-Stewart, P. D. & Blow, D. M. Microbatch crystallization under oil—a new technique allowing many small-volume crystallization trials. *J. Cryst. Growth* **122**, 176–180 (1992).
42. Otwinowski, Z. & Minor, W. in *Macromolecular Crystallography Part A* (eds Carter, C. W. Jr & Sweet, R. M.) 307–326 (Academic, New York, 1997).
43. Grosse-Kunstleve, R. W. & Adams, P. D. Substructure search procedures for macromolecular structures. *Acta Crystallogr. D* **59**, 1974–1977 (2003).
44. Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallogr. D* **59**, 2023–2030 (2003).
45. Perrakis, A., Morris, R. M. & Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458–463 (1999).
46. Terwilliger, T. SOLVE and RESOLVE: automated structure solution, density modification, and model building. *J. Synch. Radiat.* **11**, 49–52 (2004).
47. McRee, D. E. XtalView Xfit—A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165 (1999).
48. Winn, M. D., Isupov, M. N. & Murshudov, G. N. Use of TLS parameters to

model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D* **57**, 122–133 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Karniol, S. Beale and K. Satyshur for technical advice and acknowledge the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor for the support of this research program. Use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences. This work was supported by grants from the US National Science Foundation (R.D.V. and K.T.F.), the US Department of Energy (R.D.V.), and the W.M. Keck Foundation (K.T.F.).

Author Contributions J.R.W. purified protein and grew and characterized crystals; J.S.B. collected data and carried out initial phase determination; K.T.F. and J.R.W. phased, modelled, refined and validated structure; R.D.V. initiated collaboration; K.T.F. designed structure experiments; J.R.W., K.T.F. and R.D.V. interpreted the structure and prepared the manuscript and figures.

Author Information Atomic coordinates and structure factor amplitudes have been deposited in the Protein Data Bank (accession code 1ZTU). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.T.F. (forest@bact.wisc.edu).

LETTERS

The formation of stars by gravitational collapse rather than competitive accretion

Mark R. Krumholz¹, Christopher F. McKee^{2,3} & Richard I. Klein^{3,4}

There are two dominant models of how stars form. Under gravitational collapse, star-forming molecular clumps, of typically hundreds to thousands of solar masses ($M_{\odot}$), fragment into gaseous cores that subsequently collapse to make individual stars or small multiple systems^{1–3}. In contrast, competitive accretion theory suggests that at birth all stars are much smaller than the typical stellar mass ($\sim 0.5M_{\odot}$), and that final stellar masses are determined by the subsequent accretion of unbound gas from the clump^{4–8}. Competitive accretion models interpret brown dwarfs and free-floating planets as protostars ejected from star-forming clumps before they have accreted much mass; key predictions of this model are that such objects should lack disks, have high velocity dispersions, form more frequently in denser clumps^{9–11}, and that the mean stellar mass should vary within the Galaxy⁸. Here we derive the rate of competitive accretion as a function of the star-forming environment, based partly on simulation¹², and determine in what types of environments competitive accretion can occur. We show that no observed star-forming region can undergo significant competitive accretion, and that the simulations that show competitive accretion do so because the assumed properties differ from those determined by observation. Our result shows that stars form by gravitational collapse, and explains why observations have failed to confirm predictions of the competitive accretion model.

In both theories, a star initially forms when a gravitationally bound gas core collapses. The crucial distinction between them is their prediction for what happens subsequently. In gravitational collapse, after a protostar has consumed or expelled all the gas in its initial core, it may continue accreting from its parent clump. However, it will not accrete enough to change its mass substantially^{13,14}. In contrast, competitive accretion requires that the amount accreted after the initial core is consumed be substantially larger than the protostellar mass. We define $f_m \equiv \dot{m} \cdot t_{\text{dyn}} / m_*$ as the fractional change in mass that a protostar of mass m_* undergoes each dynamical time t_{dyn} of its parent clump, starting after the initial core has been consumed. Gravitational collapse holds that $f_m \ll 1$, whereas competitive accretion requires $f_m \gg 1$.

We consider a protostar embedded in a molecular clump of mass M and mass-weighted one-dimensional velocity dispersion σ . Competitive accretion theories usually begin with seed protostars of mass $m_* \approx 0.1M_{\odot}$ (refs 4–7), so we adopt this as a typical value. We consider two possible geometries: spherical clumps of radius R and filaments of radius R and length L , where $L \gg R$. These extremes bracket real star-forming clumps, which have a range of aspect ratios. The virial mass for (spherical, filamentary) clumps is:

$$M_{\text{vir}} \equiv \left(\frac{5\sigma^2 R}{G}, \frac{2\sigma^2 L}{G} \right) \quad (1)$$

and the virial parameter is $\alpha_{\text{vir}} \equiv M_{\text{vir}}/M$ (refs 15, 16). The dynamical time is $t_{\text{dyn}} \equiv R/\sigma$.

First, we suppose that the gas that the protostar is accreting is not collected into bound structures on scales smaller than the entire clump. Because the gas is unbound, we may neglect its self-gravity and treat this as a problem of a non-self-gravitating gas accreting onto a point particle. This process is Bondi–Hoyle accretion in a turbulent medium, which gives an accretion rate¹²:

$$\dot{m}_* \approx 4\pi\phi_{\text{BH}}\bar{\rho}\frac{(Gm_*)^2}{(\sqrt{3}\sigma)^3} \quad (2)$$

where $\bar{\rho}$ is the mean density in the clump. The factor ϕ_{BH} represents the effects of turbulence, which we estimate in terms of parameters σ , m_* and R in the Supplementary Information¹². From equation (2) and the definitions of the virial parameter and the dynamical time, we find that accretion of unbound gas gives:

$$f_{\text{m-BH}} = \left(14.4, 3.08 \frac{L}{R} \right) \phi_{\text{BH}} \alpha_{\text{vir}}^{-2} \left(\frac{m_*}{M} \right) \quad (3)$$

for a (spherical, filamentary) star-forming region. From this result, we can immediately see that competitive accretion is most effective in low-mass clumps with virial parameters much smaller than unity.

Tables 1 and 2 show a broad sample of observed star-forming regions. None of them have a value of $f_{\text{m-BH}}$ near unity, which is inconsistent with competitive accretion and in agreement with gravitational collapse. We note that the Bondi–Hoyle rate is an upper limit on the accretion. If the stars are sufficiently close-packed, their tidal radii will be smaller than their Bondi–Hoyle radii, and the accretion rate will be lower⁵. Also, radiation pressure will halt Bondi–Hoyle accretion onto stars larger than $\sim 10M_{\odot}$ (ref. 17).

The second possible way that a star could gain mass is by capturing and accreting other gravitationally bound cores. We can analyse this

Table 1 | Sample star-forming regions

Name	Ref.	Mass	Type	$M (M_{\odot})$	R (pc)	L (pc)	σ (km s ⁻¹)
L1495 I	30	Low	Sph.	410	2.1	–	0.58
L1495 II	30	Low	Sph.	950	2.4	–	0.67
L1709	16	Low	Fil.	140	0.23	3.6	0.48
L1755	16	Low	Fil.	171	0.15	6.3	0.53
W44	23	High	Sph.	16000	0.35	–	3.9
W75(OH)	23	High	Sph.	5600	0.25	–	3.5
f-fil	16	High	Fil.	5000	0.25	13	1.41
N-fil	16	High	Fil.	16000	2.3	88	1.54

Sph., spherical; Fil., filamentary; f-fil, Orion integral filament; N-fil, Orion North filament. For L1495 I and II, the data are from the ¹²CO observations of ref. 30, and the masses are M_{CO} from ref. 30. For W44 and W75(OH), the data are the CS $J = 5 \rightarrow 4$ observations of ref. 23, and the masses are M_n from ref. 23. (Here CS is carbon sulphide, and J is the rotational quantum number of the CS molecule.) CS $J = 5 \rightarrow 4$ is a very high-density tracer, so it biases the results to small virial parameters by excluding low-density parts of the clump.

¹Astrophysics Department, Princeton University, Princeton, New Jersey 08544, USA. ²Physics Department, ³Astronomy Department, UC Berkeley, Berkeley, California 94720, USA. ⁴Lawrence Livermore National Laboratory, Livermore, California 94550, USA.

process by some simple approximations. First, when a star collides with a core it begins accreting gas from it, causing a drag force¹⁸. If drag dissipates enough energy, the two become bound. We can therefore compute a critical velocity below which any collision will lead to a capture and above which it will not. Second, cores and stars should inherit the velocity dispersion of the gas from which they form, so we assume they have maxwellian velocity distributions with dispersion σ . The true functional form may be different, but this will only affect our estimates by a factor of order unity. Third, we neglect the range of core sizes, and assume that all cores have a generic radius R_{co} and mass M_{co} . Competitive accretion requires $M_{\text{co}} \leq m_*$, so we take $M_{\text{co}} = m_*$, which gives the highest possible capture rate. Finally, we make use of an important observational result: cores within a molecular clump have roughly the same surface density as the clump itself⁹, that is, $\Sigma = M(\pi R^2, 2RL)^{-1}$ for (spherical, filamentary) clumps. This enables us to compute the escape velocity from the surface of a core in terms of the properties of the clump:

$$v_{\text{esc}} = \left[\left(10, \sqrt{8\pi \frac{L}{R}} \right) \alpha_{\text{vir}}^{-1} \sqrt{\frac{M_{\text{co}}}{M}} \right]^{1/2} \sigma \quad (4)$$

With these approximations, it is straightforward to compute the amount of mass that a protostar can expect to gain by capturing other cores. In the Supplementary Information, we show that it is:

$$f_{\text{m-cap}} = (0.42, 0.36) \phi_{\text{co}} [4 + 2u_{\text{esc}}^2 - (4 + 7.32u_{\text{esc}}^2) \exp(-1.33u_{\text{esc}}^2)] \quad (5)$$

where ϕ_{co} is the fraction of the parent clump mass that is in bound cores and $u_{\text{esc}} \equiv v_{\text{esc}}/\sigma$. Surveys generally find core mass fractions of $\phi_{\text{co}} \approx 0.1$ (refs 20–22), so we adopt this as a typical value, giving the numerical values of $f_{\text{m-cap}}$ shown in Table 2. As with $f_{\text{m-BH}}$, all the estimated values are well below unity.

If we let $f_{\text{m}} = f_{\text{m-BH}} + f_{\text{m-cap}}$, then we can use our simple models to determine where in parameter space a star-forming clump must fall to have $f_{\text{m}} \geq 1$. For simplicity, we consider a spherical clump with fixed $\phi_{\text{BH}} = 5$ and $\phi_{\text{co}} = 0.1$ (typical values for observed regions), and a seed protostar of mass $m_* = 0.1$. In this case, both $f_{\text{m-BH}}$ and $f_{\text{m-cap}}$ are functions of $\alpha_{\text{vir}}^2 M$ alone, and we find $f_{\text{m}} \geq 1$ for $\alpha_{\text{vir}}^2 M < 8.4 M_{\odot}$. The functional dependence is more complex if we include filamentary regions and allow ϕ_{BH} and ϕ_{co} to vary, but the qualitative result is unchanged. Observed star-forming regions have $\alpha_{\text{vir}} \approx 1$ and $M \approx 10^2 - 10^4 M_{\odot}$ (ref. 23), which produces $f_{\text{m}} \ll 1$. No known star-forming region has $\alpha_{\text{vir}}^2 M$ small enough for competitive accretion to work. Thus, the cores from which stars form must contain all the mass they will ever have, which is the gravitational collapse model.

Our simple estimate of f_{m} is consistent with simulations of competitive accretion as well, and explains why competitive accretion works in the simulations. All competitive accretion simulations have virial parameters $\alpha_{\text{vir}} \ll 1$. In some cases the simulations start in this condition^{5,6,24,25}, with $\alpha_{\text{vir}} \approx 0.01$ as a typical choice. In other cases, the virial parameter is initially of order unity, but as turbulence decays in the simulation it decreases to $\ll 1$ in roughly a crossing time^{7,9,10,26}. Once competitive accretion gets going, these simulations reach $\alpha_{\text{vir}} \ll 1$ as well. In addition, many of the simulations consider

star-forming clumps of masses considerably smaller than the $\sim 5,000 M_{\odot}$ typical of most galactic star formation²³, with $M \lesssim 100 M_{\odot}$ not uncommon. Consequently, the simulations have $\alpha_{\text{vir}}^2 M \lesssim 10 M_{\odot}$, which explains why they find competitive accretion to be important. Note that simulations where turbulence decays will have $\phi_{\text{BH}} \approx 1$, rather than the typical value of $\phi_{\text{BH}} = 5$ we have used for real regions, but this does not substantially modify our conclusions.

Three other aspects of the simulations increase even further their estimate of f_{m} . First, the Bondi–Hoyle radius of a $0.1 M_{\odot}$ seed protostar in a typical clump is only 5 AU (astronomical units), a smaller scale than any of the competitive accretion simulations resolve. This under-resolution may enhance accretion¹². Second, small virial parameters lead most of the mass to collapse to stars, giving $\phi_{\text{co}} \approx 0.5$ –1 after a dynamical time, and also tend to make the cloud fragment into smaller pieces, lowering M . Third, rapid collapse leaves no time for large cores to assemble. For example, one simulation of a $\sim 1,000 M_{\odot}$ clump produces no cores larger than $1 M_{\odot}$ (ref. 7), inconsistent with observations that find numerous cores more massive than this in similar regions^{22,27}. With no large cores, large stars can form only via competitive accretion.

Thus, our results are consistent with the simulations, but they show that the simulations are not modelling realistic star-forming clumps. One might argue that all clumps do enter a phase with $\alpha_{\text{vir}} \ll 1$ that occurs rapidly and has therefore never been observed, but that most stars are formed during this collapse phase. In this scenario, though, protostars associated with observed star-forming regions should have systematically lower masses than the field star population, because they were formed before the collapse phase in which competitive accretion might occur. We would expect to see a systematic variation in mean stellar mass with age in young clusters, corresponding to cluster evolution into a state more and more favourable to competitive accretion. We do not observe this.

We hypothesize that the primary problem with the simulations—the reason they evolve to $\alpha_{\text{vir}} \ll 1$ —is that they omit feedback from star formation. Recent observations of protostellar outflow cavities show that outflows inject enough energy to sustain the turbulence and prevent the virial parameter from declining to values much less than unity²⁸. Another possible problem in the simulations is that they simulate isolated clumps containing too little material. Real clumps are embedded in molecular clouds, and large-scale turbulent motions in the clouds may cascade down to the clump scale and prevent the turbulence from decaying. A third possibility is that turbulence decays too quickly in the simulations because they do not include magnetic fields and their initial velocity fields, unlike in real clumps, are balanced rather than imbalanced between left- and right-propagating modes²⁹.

One implication of our work is that brown dwarfs need not have been ejected from their natal clump, so their velocity dispersions should be at most slightly greater than those of stars, and their frequency need not change as a function of clump density. This also removes a discrepancy between observations showing that brown dwarfs have disks¹¹ and theoretical models of their origins. We also conclude that the mean stellar mass need not vary from one star-forming region to another as competitive accretion predicts, removing a discrepancy between theory⁸ and observations that have thus far failed to find any substantial variation in typical stellar mass with the star-forming environment. In the gravitational collapse scenario, the mean stellar mass may be roughly constant in the Galaxy, but may vary with the background radiation field in starburst regions and in the early Universe³.

Received 12 August; accepted 29 September 2005.

- Shu, F. H., Adams, F. C. & Lizano, S. Star formation in molecular clouds—observation and theory. *Annu. Rev. Astron. Astrophys.* **25**, 23–81 (1987).
- Padoan, P. & Nordlund, Å. The stellar initial mass function from turbulent fragmentation. *Astrophys. J.* **576**, 870–879 (2002).

Table 2 | Computed properties of sample star-forming regions

Name	α_{vir}	ϕ_{BH}^{12}	$f_{\text{m-BH}}$	u_{esc}	$f_{\text{m-cap}}$
L1495 I	2.0	2.4	0.0022	0.28	0.0015
L1495 II	1.3	3.0	0.0026	0.28	0.0015
L1709	2.8	0.93	0.0042	0.44	0.0072
L1755	4.8	0.54	0.0017	0.40	0.0052
W44	0.39	6.4	0.0038	0.25	0.0010
W75(OH)	0.63	5.2	0.0034	0.26	0.0011
f-fil	2.4	4.1	0.0023	0.26	0.00097
N-fil	6.2	5.7	0.0001	0.11	0.00003

3. Larson, R. B. Thermal physics, cloud geometry and the stellar initial mass function. *Mon. Not. R. Astron. Soc.* **359**, 211–222 (2005).
4. Bonnell, I. A., Bate, M. R. & Zinnecker, H. On the formation of massive stars. *Mon. Not. R. Astron. Soc.* **298**, 93–102 (1998).
5. Bonnell, I. A., Bate, M. R., Clarke, C. J. & Pringle, J. E. Competitive accretion in embedded stellar clusters. *Mon. Not. R. Astron. Soc.* **323**, 785–794 (2001).
6. Bonnell, I. A., Clarke, C. J., Bate, M. R. & Pringle, J. E. Accretion in stellar clusters and the initial mass function. *Mon. Not. R. Astron. Soc.* **324**, 573–579 (2001).
7. Bonnell, I. A., Vine, S. G. & Bate, M. R. Massive star formation: nurture, not nature. *Mon. Not. R. Astron. Soc.* **349**, 735–741 (2004).
8. Bate, M. R. & Bonnell, I. A. The origin of the initial mass function and its dependence on the mean Jeans mass in molecular clouds. *Mon. Not. R. Astron. Soc.* **356**, 1201–1221 (2005).
9. Bate, M. R., Bonnell, I. A. & Bromm, V. The formation mechanism of brown dwarfs. *Mon. Not. R. Astron. Soc.* **332**, L65–L68 (2002).
10. Bate, M. R., Bonnell, I. A. & Bromm, V. The formation of a star cluster: predicting the properties of stars and brown dwarfs. *Mon. Not. R. Astron. Soc.* **339**, 577–599 (2003).
11. Mohanty, S., Jayawardhana, R. & Basri, G. The T Tauri phase down to nearly planetary masses: echelle spectra of 82 very low mass stars and brown dwarfs. *Astrophys. J.* **626**, 498–522 (2005).
12. Krumholz, M. R., McKee, C. F. & Klein, R. I. Bondi-Hoyle accretion in a turbulent medium. *Astrophys. J.* (in the press); preprint at (<http://arXiv.org/astro-ph/0510410>) (2005).
13. McKee, C. F. & Tan, J. C. The formation of massive stars from turbulent cores. *Astrophys. J.* **585**, 850–871 (2003).
14. Padoan, P., Kritsuk, A., Norman, M. L. & Nordlund, Å. A solution to the pre-main-sequence accretion problem. *Astrophys. J. Lett.* **622**, L61–L64 (2005).
15. Bertoldi, F. & McKee, C. F. Pressure-confined clumps in magnetized molecular clouds. *Astrophys. J.* **395**, 140–157 (1992).
16. Fiege, J. D. & Pudritz, R. E. Helical fields and filamentary molecular clouds—I. *Mon. Not. R. Astron. Soc.* **311**, 85–104 (2000).
17. Edgar, R. & Clarke, C. The effect of radiative feedback on Bondi-Hoyle flow around a massive star. *Mon. Not. R. Astron. Soc.* **349**, 678–686 (2004).
18. Ruffert, M. & Arnett, D. Three-dimensional hydrodynamic Bondi-Hoyle accretion. 2: Homogeneous medium at mach 3 with $\gamma = 5/3$. *Astrophys. J.* **427**, 351–376 (1994).
19. Larson, R. B. Turbulence and star formation in molecular clouds. *Mon. Not. R. Astron. Soc.* **194**, 809–826 (1981).
20. Motte, F., Andre, P. & Neri, R. The initial conditions of star formation in the ρ Ophiuchi main cloud: wide-field millimeter continuum mapping. *Astron. Astrophys.* **336**, 150–172 (1998).
21. Testi, L. & Sargent, A. I. Star formation in clusters: A survey of compact millimeter-wave sources in the Serpens core. *Astrophys. J. Lett.* **508**, L91–L94 (1998).
22. Johnstone, D., Fich, M., Mitchell, G. F. & Moriarty-Schieven, G. Large area mapping at 850 microns. III. Analysis of the clump distribution in the Orion B molecular cloud. *Astrophys. J.* **559**, 307–317 (2001).
23. Plume, R., Jaffe, D. T., Evans, N. J., Martin-Pintado, J. & Gomez-Gonzalez, J. Dense gas and star formation: Characteristics of cloud cores associated with water masers. *Astrophys. J.* **476**, 730–749 (1997).
24. Klessen, R. S. & Burkert, A. The formation of stellar clusters: gaussian cloud conditions I. *Astrophys. J. Suppl.* **128**, 287–319 (2000).
25. Klessen, R. S. & Burkert, A. The formation of stellar clusters: gaussian cloud conditions II. *Astrophys. J.* **549**, 386–401 (2001).
26. Bate, M. R., Bonnell, I. A. & Bromm, V. The formation of close binary systems by dynamical interactions and orbital decay. *Mon. Not. R. Astron. Soc.* **336**, 705–713 (2002).
27. Beuther, H. & Schilke, P. Fragmentation in massive star formation. *Science* **303**, 1167–1169 (2004).
28. Quillen, A. C. *et al.* Turbulence driven by outflow-blown cavities in the molecular cloud of NGC 1333. *Astrophys. J.* (in the press); preprint at (<http://arXiv.org/astro-ph/0503167>) (2005).
29. Cho, J. & Lazarian, A. Compressible magnetohydrodynamic turbulence: mode coupling, scaling relations, anisotropy, viscosity-damped regime and astrophysical implications. *Mon. Not. R. Astron. Soc.* **345**, 325–339 (2003).
30. Kramer, C. & Winnewisser, G. A molecular survey of the dark cloud L 1495 in Taurus. *Astron. Astrophys. Suppl.* **89**, 421–428 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. T. Fisher for discussions and P. Padoan for comments. This work was supported by grants from NASA through the Hubble Fellowship, GSRP and ATP programmes, by the NSF, and by the US DOE through the Lawrence Livermore National Laboratory. Computer simulations for this work were performed at the San Diego Supercomputer Center (supported by the NSF), the National Energy Research Scientific Computer Center (supported by the US DOE), and Lawrence Livermore National Laboratory (supported by the US DOE). M.R.K. is a Hubble Fellow.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.R.K. (krumholz@astro.princeton.edu).

Nanofabricated media with negative permeability at visible frequencies

A. N. Grigorenko¹, A. K. Geim¹, H. F. Gleeson¹, Y. Zhang¹, A. A. Firsov^{1,2}, I. Y. Khrushchev³ & J. Petrovic³

A great deal of attention has recently been focused on a new class of smart materials—so-called left-handed media—that exhibit highly unusual electromagnetic properties and promise new device applications^{1–6}. Left-handed materials require negative permeability μ , an extreme condition that has so far been achieved only for frequencies in the microwave to terahertz range^{7–11}. Extension of the approach described in ref. 7 to achieve the necessary high-frequency magnetic response in visible optics presents a formidable challenge^{12–15}, as no material—natural or artificial—is known to exhibit any magnetism at these frequencies. Here we report a nanofabricated medium consisting of electromagnetically coupled pairs of gold dots with geometry carefully designed at a 10-nm level. The medium exhibits a strong magnetic response at visible-light frequencies, including a band with negative μ . The magnetism arises owing to the excitation of an antisymmetric plasmon resonance. The high-frequency permeability qualitatively reveals itself via optical impedance matching. Our results demonstrate the feasibility of engineering magnetism at visible frequencies and pave the way towards magnetic and left-handed components for visible optics.

Landau and Lifshitz argued¹² “there is certainly no meaning in using the magnetic susceptibility from optical frequencies onwards, and in discussion of such phenomena we must put $\mu = 1$ ”. This statement is strongly supported by experiment: the magnetic susceptibility χ of all natural materials tails off at microwave frequencies. Still, there may be a way to overcome the fundamental limitations, as shown by Pendry *et al.* who have suggested exploiting the inductive response from structured non-magnetic materials to obtain high-frequency magnetism⁷. The idea was successfully implemented by using arrays of copper split-rings that generated a magnetic response at frequencies up to 100 THz (refs 8–11). In this case, magnetic properties emerge owing to collective motion of a large number of electrons, and theoretical arguments (as, for example, in ref. 12) valid for individual electrons and atoms no longer hold. It is tempting to extend the approach further to visible-light frequencies, where one can expect most applications. However, the direct scaling of the demonstrated microwave media to visible optics is problematic. This would require split-ring-like structures with sizes down to 100 nm and critical features^{9–11} controlled on the level of ~ 10 nm, which is technologically difficult to achieve. More importantly, the scaling could fail in principle because of different electromagnetic responses of materials to visible light and microwaves (for example, it was predicted that inherent losses should limit the approach demonstrated in refs 7–11 to frequencies well below optical^{9,13}).

In this work, by employing a novel geometry, we make a critical step of demonstrating metamaterials with magnetic response at frequencies in the visible spectrum. The design of our media follows recent theoretical suggestions^{14,15}, and relies on antisymmetric plasmon resonances in a simple pair of short metal pillars of a

submicrometre size (Fig. 1). The simplification of the resonator geometry with respect to the double split-rings geometry used in microwaves is important for two reasons. First, this allows the use of current lithography techniques to fabricate metallic structures with plasmon resonances at visible-light frequencies. Second, the simple geometry reduces the number of resonant modes interacting with a light field, which consequently leads to a reduction in energy losses. Note that such changes in geometry also reflect the trend known from lasing techniques, where complex closed resonators used in masers were replaced by simple open resonators (pair of mirrors) in lasers^{16,17}.

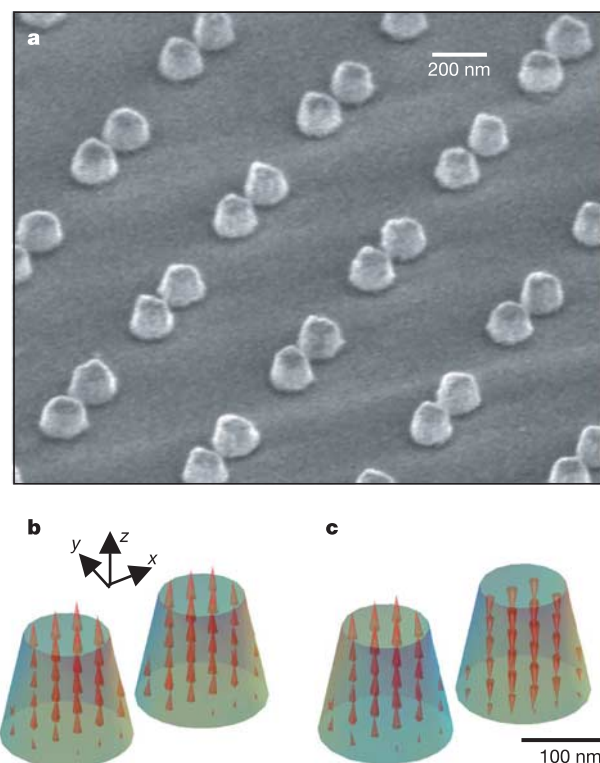


Figure 1 | Nanofabricated medium with magnetic response at optical frequencies. **a**, Scanning electron micrograph (viewed at an angle) of an array of Au nanopillars. **b**, **c**, Numerical simulation of the distribution of electric currents (arrows) inside a pair of such pillars for the symmetric and antisymmetric resonant z-modes, respectively. The non-cylindrical shape of pillars is important to provide an efficient coupling to incident light, and was intentionally introduced in our design through a choice of microfabrication procedures.

¹Department of Physics and Astronomy, University of Manchester, Manchester, M13 9PL, UK. ²Institute of Microelectronics Technology, 142432 Chernogolovka, Russia.

³Department of Electronic Engineering, Aston University, Aston Triangle, Birmingham B4 7ET, UK.

Figure 1 shows an example of our devices and illustrates the basic idea behind the experiment. The prepared structures were large arrays of Au pillars fabricated by high-resolution electron-beam lithography on a glass substrate and grouped in tightly spaced pairs (except for reference samples consisting of similar but isolated Au pillars). The structures typically covered an area of $\sim 0.1 \text{ mm}^2$ and contained $\sim 10^6$ pillars. The lattice constant, a , for periodic arrays was down to 400 nm—that is, smaller than the wavelength λ of visible light. Heights h of Au pillars (80–90 nm) and their diameters $d \approx 100 \text{ nm}$ were chosen through numerical simulations so that the plasmon resonance in the reference samples appeared at red-light wavelengths, $\lambda \approx 670 \text{ nm}$. A number of different structures were studied with d between 80 nm and 140 nm and the pair separation s between centres of adjacent pillars in the range 140 nm to 200 nm; that is, the gap $s - d$ between the neighbouring pillars varied from 100 nm down to almost zero. (The best results were achieved at $s = 200 \text{ nm}$, $d = 140 \text{ nm}$, $h = 80 \text{ nm}$, and $s = 140 \text{ nm}$, $d = 110 \text{ nm}$, $h = 90 \text{ nm}$.) At these separations, electromagnetic interaction between neighbouring pillars within a pair is important and plasmon resonance observed for an individual pillar splits into two resonances for a pillar pair. These resonances are referred to as symmetric and antisymmetric, similar to the case of any classical or quantum system with two interacting parts and in agreement with the notation used for plasmon resonances in nanoparticles¹⁸.

There exist three symmetric and three antisymmetric main resonant modes in an interacting pair with currents flowing along the x , y and z axes. Figure 1 shows the symmetric (Fig. 1b) and the antisymmetric (Fig. 1c) z -modes calculated for our experimental geometry using Femlab software (Comsol Inc.). For the symmetric resonance, electrons in neighbouring pillars move in phase and generate overall a dipole contribution to permittivity ϵ , similar to isolated or non-interacting pillars. In the antisymmetric z -mode, however, electrons move in anti-phase so that the oscillating dipoles cancel each other leaving only magnetodipole and quadrupole responses¹⁹. One can see in Fig. 1c that the anti-phase movement of electrons along the z axis effectively results in a current loop in the z - x plane (note that a pair of pillars can be thought of as a ring with two slits at the opposite sides). This high-frequency electric current generates a magnetic moment in the y -direction and contributes to permeability μ . Not all resonant modes are necessarily excited by incident light. Modes' coupling to light is governed by the symmetry group of a pillar pair, which is C_{2v} in our case. For this group, normal incident light with the electric field along the x axis (later referred to as TM polarization) is coupled to both the dipole symmetric x -mode and the magnetic antisymmetric z -mode, while normal light with the electric field along the y axis (TE polarization) excites only the dipole symmetric y -mode, as discussed in ref. 20. The choice of the symmetry group is important for creating magnetic response (for example, magnetic modes are not necessarily excited in a pair of infinite cylinders with the symmetry group C_{2h} ; ref. 21).

The symmetry of our pillar pairs thus implies that an array of pillar pairs should exhibit two main plasmon resonances (z -antisymmetric and x -symmetric) for TM light of normal incidence, one of which (z -antisymmetric) disappears and the other (x -symmetric) changes into y -symmetric resonance as we rotate polarization by 90° , to TE polarization. On the other hand, an array of isolated pillars should demonstrate only one resonance for both polarizations. Our experiments confirmed this. Figure 2 shows the reflection spectra measured in normal incident light of TM and TE polarizations (green and red curves, respectively) for arrays made of pillar pairs of the same dimensions but different lattice constants, as well as for the reference array of isolated pillars (micrographs of the corresponding samples are shown next to their spectra.) The reference array exhibits only one resonance at $\lambda \approx 670 \text{ nm}$ for both polarizations (Fig. 2g and h), in agreement with our symmetry and numerical analyses. On the other hand, arrays made of the same pillars but packed in interacting pairs showed two distinct resonances in the TM spectra (Fig. 2a, c and e). If

the polarization of incident light was rotated by 90° , the high-frequency, 'green' resonance disappeared, and the reflection spectra showed only the low-frequency, 'red' resonance, albeit with its position slightly shifted (Fig. 2b, d and f). The change in the reflection spectra was so marked that the colour of the samples viewed in white light changed (see photographs in Fig. 2). Our structures looked amber for TE light, similar in colour to the media consisting of isolated pillars. In stark contrast, for TM polarization, the structures look green owing to a large contribution from the green resonance.

The spectral positions of red and green resonance peaks did not depend on the lattice constant for all studied samples with pillar pairs of the same dimensions (Fig. 2c and e shows two examples with $a = 400$ and 600 nm). This excludes diffraction as a possible origin for the observed resonances. This was further confirmed by making random arrays of pillar pairs, which did not influence the spectral positions of the resonance peaks. (It is worth noting that, contrary to our case, diffraction modes play a prominent role in light reflection from perforated metal²² or arrays of nanoparticles deposited on metallic substrates²³.) Although the resonance positions did not depend on a , they were strongly affected by changing separation s and by covering the structure with a dielectric medium, as expected for resonances split by electromagnetic interaction between pillars. As an example, Fig. 2a and b shows reflection spectra of the same sample as in Fig. 2c and d but covered by an optically thin ($< 30 \text{ nm}$) layer of glycerine. This resulted in a notable red-shift of the green resonance peak by $\delta\lambda \approx 50 \text{ nm}$, in agreement with red-shifts

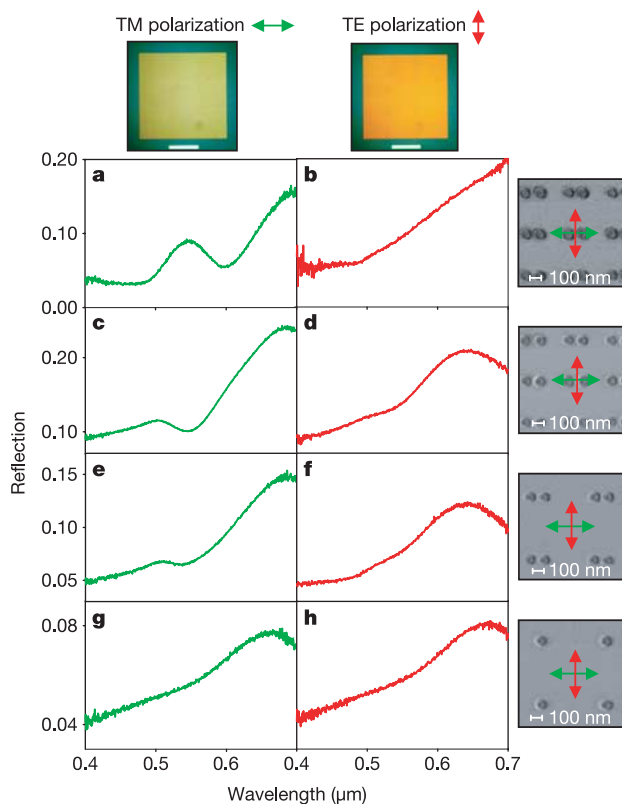


Figure 2 | Experimental reflection spectra for our nanostructured media. Green and red curves are for TM and TE polarizations of normal incident light, respectively. Micrographs of the studied samples are shown on the right. For all the samples, pillars have the same separation $s = 140 \text{ nm}$, height $h = 90 \text{ nm}$ and average diameter $d = 110 \text{ nm}$. Spectra **a**, **b**, are for the sample of **c**, **d**, but covered with an optically thin layer of glycerine; for **c**, **d**, the lattice constant $a = 400 \text{ nm}$; for **e**, **f**, $a = 600 \text{ nm}$; **g**, **h**, are for isolated pillars with $a = 600 \text{ nm}$. The top photographs show images of the sample **a**, **b**, in white light for two polarizations.

reported for multipole resonances in nanoparticles²⁴. Moreover, we observed a significant increase (by a factor of 3) in the strength of the green resonance. Note that, in this case, the structure reflected as much as 10% of light near the green resonance peak (Fig. 2a), which is a significant magnitude under our experimental conditions (for comparison, a glass substrate with $\varepsilon = 2.25$ reflects only 4% of the incident light).

The described experiments assign the green-light resonance (disappearing in TE polarization) unambiguously to the antisymmetric z -mode, which gives rise to μ , and the red-light resonances to the x - and y -symmetric modes, contributing to ε (refs 14, 15). To further strengthen this identification of the resonances, we compare the measured spectra with theory (Fig. 3). The theoretical calculations were performed with the Electromagnetic module of Femlab software, which solves Maxwell's equations for the actual experimental geometry. We described pillars by Drude's conductivity

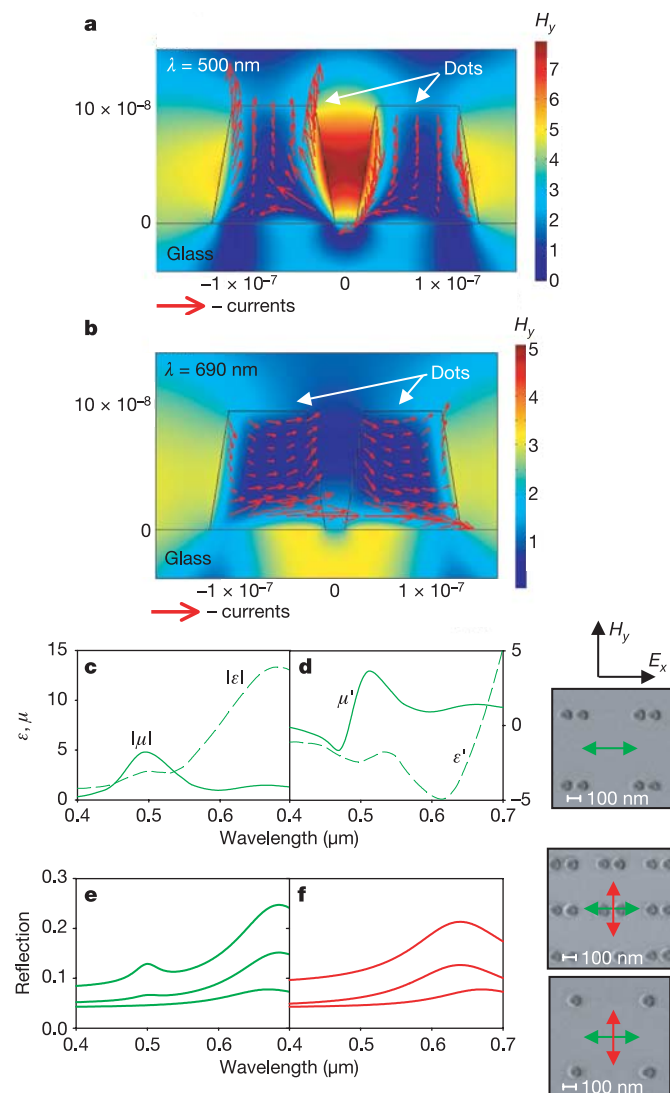


Figure 3 | Numerical simulations of optical response for interacting Au pillars. **a, b**, Distribution of electric currents (red arrows) and magnetic field H_y (colour map measured in units of the magnetic field amplitude of the incident wave) for the pillars being illuminated by normal incident light of TM polarization with wavelengths $\lambda = 500$ nm (**a**) and $\lambda = 690$ nm (**b**). Geometrical sizes are shown in metres at the bottom and the left of images. 'Dots' in the image refer to Au nanopillars. **c, d**, The spectral dependence of permittivity ε and permeability μ : **c** shows the absolute values and **d** the real parts. **e, f**, Calculated reflection spectra corresponding to the experimental situation in Fig. 2.

$\sigma(\omega) = \sigma_0 / (1 - i\omega\tau)$, where ω is the angular frequency of light, $\sigma_0 = 3.7 \times 10^{17} \text{ s}^{-1}$ the conductivity of gold and $\tau = 2.4 \times 10^{-14} \text{ s}$ the scattering time, and the glass substrate by a dielectric with permittivity $\varepsilon = 2.25$. The calculations have confirmed that TM green light of normal incidence with $\lambda = 500$ nm excites anti-phase currents flowing mostly along the z axis (Fig. 3a), which is a characteristic of the antisymmetric z -mode. A current loop is effectively created in the z - x plane, which generates strong magnetic fields H_y in the region between pillars (the red region of Fig. 3a) and produces a substantial magnetic moment. On the other hand, TM red light at $\lambda = 690$ nm excites mostly in-phase currents flowing along the x axis, which do not produce any significant magnetic moment (Fig. 3b).

We have also calculated the spectral dependence of effective permeability and permittivity of our structures. Figure 3 shows the absolute values and real parts for ε_x and μ_y (denoted here as ε and μ) for the tightest pair packing (the curves were obtained by solving Maxwell's equations for an electromagnetic wave interacting with a periodic array on top of a glass substrate). The calculated spectral positions of the peaks in ε and μ are in good agreement with the peak positions found experimentally, and the theoretical reflection spectra also agree well with the experimental data. The calculations show that our microstructures have considerable permeability $\mu' \equiv \text{Re}(\mu)$ ranging from -1 to 3 for the array of tightest packing (Fig. 3d). For the sample of Fig. 2c, for example, its calculated permeability varied between 0.5 and 1.5 , and for Fig. 2a μ' varied between -0.25 and 2.1 .

As a complementary analysis to the numerical simulations, we have employed another approach that is routinely used in optics to extract material parameters from reflection spectra^{25,26}. To this end, we noted that, neglecting mode interaction, the calculated resonances in permittivity followed the standard dispersion relation²⁵ $\Delta\varepsilon(\lambda) = f_s \lambda^2 / (\lambda^2 - \lambda_s^2 - i\lambda\Delta\lambda_s)$, where λ_s is the wavelength of the symmetric resonance, $\Delta\lambda_s$ its half-width and f_s the effective oscillator

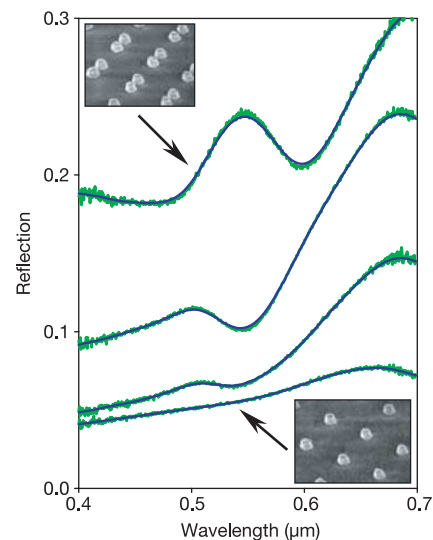


Figure 4 | Example of fitting the experimental reflection spectra with theory. The green curves are the measured spectra of Fig. 2; the blue curves show the best fit based on the dispersion relations described in the main text. The parameters of the main resonances extracted from the fitting curves are as follows. Single pillars; $\lambda_s \approx 665$ nm, $\Delta\lambda_s \approx 165$ nm and $f_s \approx 0.79$. Pairs with $a = 600$ nm; $\lambda_s \approx 690$ nm, $\Delta\lambda_s \approx 147$ nm, $f_s \approx 1.76$ and $\lambda_a \approx 550$ nm, $\Delta\lambda_a \approx 85$ nm, $f_a \approx 0.06$. Pairs with $a = 400$ nm; $\lambda_s \approx 685$ nm, $\Delta\lambda_s \approx 147$ nm, $f_s \approx 3.27$ and $\lambda_a \approx 552$ nm, $\Delta\lambda_a \approx 80$ nm, $f_a \approx 0.16$. Pairs covered with a glycerine film; $\lambda_s \approx 710$ nm, $\Delta\lambda_s \approx 140$ nm, $f_s \approx 1.7$ and $\lambda_a \approx 598$ nm, $\Delta\lambda_a \approx 80$ nm, $f_a \approx 0.32$. The spectrum of the sample covered with a glycerine film is offset for clarity. The insets show micrographs of the corresponding samples.

strength. The spectral dependence of calculated permeability is also described well by the 'Pendry-type' expression^{7,10} $\mu(\lambda) = 1 + f_a \lambda_a^2 / (\lambda^2 - \lambda_a^2 - i\lambda \Delta \lambda_a)$, where λ_a , $\Delta \lambda_a$ and f_a are the same notations as above but for the antisymmetric resonance (see Supplementary Information for details). The combination of these dispersion relations with the Fresnel's reflection coefficients for a thin anisotropic film placed on a glass substrate^{25,26} leads to relatively simple expressions that can be used to fit the experimental spectra and find ϵ and μ . For example, in the case of Fig. 2c, the best fit to its spectra yields $\chi' \approx -0.5$ ($\mu' \approx 0.5$) near $\lambda = 470$ nm and corroborates the results of the Femlab calculations (see Fig. 4). Similarly, the spectra for the samples covered with a glycerine film and showing an increased strength in the antisymmetric resonance yielded a significant increase in $|\chi|$, such that negative values of μ' were achieved almost routinely. For example, χ' was about -1.3 at the green resonance in Fig. 2a (that is, $\mu' \approx -0.3$). Although our structures exhibited both negative μ' and negative ϵ' within the same range of λ (for example, $\epsilon' \approx -0.7$ and $\mu' \approx -0.3$ at the green resonance in Fig. 2a), μ had a rather large imaginary component ($\mu'' \equiv \text{Im}(\mu) \approx 1i$ at the resonance), which so far has not allowed the observation of negative refraction.

We have however observed another effect—optical impedance matching—that is more tolerant to dissipation but also exclusive to materials with a finite permeability. The effect of impedance matching is characterized by the total suppression of reflection from an interface between two media with different refraction indices, $n = (\epsilon\mu)^{1/2}$, but the same impedance values, $Z = (\epsilon/\mu)^{1/2}$. This impedance matching is well known for electrical and microwave circuits but was never observed for conventional optics, because it requires $\mu \neq 1$ (except for singular cases of optical waveguides²⁷), which has been unachievable until now. In our case, this phenomenon resulted in a total invisibility of our structured films at green-resonance frequencies for TM polarization of incident light, while the films could still be seen by using phase contrast imaging or TE polarization. For brevity, the experiments are described in Supplementary Information.

Received 19 May; accepted 19 September 2005.

- Veselago, V. G. The electrodynamics of substances with simultaneously negative values of permittivity and permeability. *Sov. Phys. Usp.* **10**, 509–514 (1968).
- Pendry, J. B. Negative refraction makes a perfect lens. *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
- Shelby, R. A., Smith, D. R. & Schultz, S. Experimental verification of a negative index of refraction. *Science* **292**, 77–79 (2001).
- Smith, D. R., Padilla, W. J., Vier, D. C., Nemat-Nasser, S. C. & Schultz, S. Composite medium with simultaneously negative permeability and permittivity. *Phys. Rev. Lett.* **84**, 4184–4187 (2000).
- Pendry, J. B. Positively negative. *Nature* **423**, 22–23 (2003).
- Houck, A. A., Brock, J. B. & Chuang, I. L. Experimental observations of a left-handed material that obeys Snell's law. *Phys. Rev. Lett.* **90**, 137401 (2003).
- Pendry, J. B., Holden, A. J., Robbins, D. J. & Stewart, W. J. Magnetism from conductors and enhanced nonlinear phenomena. *IEEE Trans. Microwave Theory Tech.* **47**, 2075–2084 (1999).
- Wiltshire, M. C. K. *et al.* Microstructured magnetic materials for RF flux guides in magnetic resonance imaging. *Science* **291**, 849–851 (2001).
- Pendry, J. B. & O'Brien, S. Magnetic activity at infrared frequencies in structured photonic crystals. *J. Phys. Condens. Matter* **14**, 6383–6394 (2002).
- Yen, T. J. *et al.* Terahertz magnetic response from artificial materials. *Science* **303**, 1494–1496 (2004).
- Linden, S. *et al.* Magnetic response of metamaterials at 100 terahertz. *Science* **306**, 1351–1353 (2004).
- Landau, L. D. & Lifshitz, E. M. *Electrodynamics of Continuous Media* Section 60 (Oxford, Pergamon, 1960).
- Dimmock, J. O. Losses in left-handed materials. *Opt. Express* **11**, 2397–2402 (2003).
- Panina, L. V., Grigorenko, A. N. & Makhnovskiy, D. P. Metal-dielectric medium with conducting nanoelements. *Phys. Rev. B* **66**, 155411 (2002).
- Podolskiy, V. A., Sarychev, A. K. & Shalaev, V. M. Plasmon modes in metal nanowires and left-handed materials. *J. Nonlinear Opt. Phys. Mater.* **11**, 65–74 (2002).
- Prokhorov, A. M. Molecular amplifier and generator for submillimeter waves. *Zh. Eksp. Teor. Fiz.* **34**, 1658–1659 (1958).
- Schawlow, A. L. & Townes, C. H. Infrared and optical masers. *Phys. Rev.* **112**, 1940–1949 (1958).
- Aizpurua, J. *et al.* Optical properties of gold nanorings. *Phys. Rev. Lett.* **90**, 057401 (2003).
- Jin, R. *et al.* Controlling anisotropic nanoparticle growth through plasmon excitation. *Nature* **425**, 487–490 (2003).
- Barron, L. D. *Molecular Light Scattering and Optical Activity* (Cambridge Univ. Press, Cambridge, UK, 1982).
- Kottmann, J. P. & Martin, O. J. F. Plasmon resonant coupling in metallic nanowires. *Opt. Express* **8**, 655–663 (2001).
- Ebbesen, T. W., Lezec, H. J., Ghaemi, H. F., Thio, T. & Wolff, P. A. Plasmon-assisted transmission of entangled photons. *Nature* **391**, 667–669 (1998).
- Felidj, N. *et al.* Enhanced substrate-induced coupling in two-dimensional gold nanoparticle arrays. *Phys. Rev. B* **66**, 245407 (2002).
- Mock, J. J., Smith, D. R. & Schultz, S. Local refractive index dependence of plasmon resonance spectra from individual nanoparticles. *Nano Lett.* **3**, 485–491 (2003).
- Born, M. & Wolf, E. *Principles of Optics* Section 2.3 (Cambridge Univ. Press, Cambridge, UK, 1980).
- Abeles, F. in *Physics of Thin Films* Vol. 6 (eds Francombe, M. H. & Hoffman, R. W.) Ch. V–VII (Academic, New York, 1971).
- Gademann, A., Durkan, C. & Shvets, I. V. Optical impedance matching with near-field optical microscopy. *J. Phys. D* **36**, 2193–2197 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by EPSRC (UK). We thank L. Panina and D. Makhnovskiy for discussions, and N. Roberts for collaboration in experiments. A.A.F. acknowledges support from INTAS.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.N.G. (sasha@man.ac.uk).

Spin-torque diode effect in magnetic tunnel junctions

A. A. Tulapurkar^{1,2,†}, Y. Suzuki^{1,2,3}, A. Fukushima^{1,2}, H. Kubota^{1,2}, H. Maehara⁴, K. Tsunekawa⁴, D. D. Djayaprawira⁴, N. Watanabe⁴ & S. Yuasa^{1,2}

There is currently much interest in the development of ‘spintronic’ devices, in which harnessing the spins of electrons (rather than just their charges) is anticipated to provide new functionalities that go beyond those possible with conventional electronic devices. One widely studied example of an effect that has its roots in the electron’s spin degree of freedom is the torque exerted by a spin-polarized electric current on the spin moment of a nanometre-scale magnet. This torque causes the magnetic moment to rotate^{1–19} at potentially useful frequencies. Here we report a very different phenomenon that is also based on the interplay between spin dynamics and spin-dependent transport, and which arises from unusual diode behaviour. We show that the application of a small radio-frequency alternating current to a nanometre-scale magnetic tunnel junction^{20–22} can generate a measurable direct-current (d.c.) voltage across the device when the frequency is resonant with the spin oscillations that arise from the spin-torque effect: at resonance (which can be tuned by an external magnetic field), the structure exhibits different resistance states depending on the direction of the current. This behaviour is markedly different from that of a conventional semiconductor diode²³, and could form the basis of a nanometre-scale radio-frequency detector in telecommunication circuits.

We performed experiments on a magnetic tunnel junction (MTJ) in the structure Si (substrate)/PtMn (15 nm)/CoFe (2.5 nm)/Ru (0.85 nm)/CoFeB (3 nm)/MgO (0.85 nm)/CoFeB (3 nm); see Fig. 1a. This multi-layered film was further patterned into oval-shaped pillars of dimension 200 nm × 100 nm, using electron-beam lithography and ion milling techniques. The bottom anti-ferromagnetically coupled CoFe and CoFeB layers (the synthetic antiferromagnetic layer) act as a pinned layer, while the top CoFeB layer acts as a free layer, whose magnetization can be changed. The resistance of the MTJ depends on the relative orientations of the pinned and free layers. The present MTJ shows a giant tunnelling magnetoresistance (TMR) due to the crystalline MgO (001) tunnelling barrier. A current passing through the MTJ gets spin-polarized by the pinned layer, and exerts a torque on the free layer.

The experimental arrangement to measure the diode effect is shown in Fig. 1a. A bias T is used to pass high-frequency current (200 MHz to 15 GHz) through the MTJ and to measure the d.c. voltage simultaneously. For all the experiments described here, the external magnetic field was applied at an angle of 30° from the pinned-layer magnetization axis within the film plane (see inset of Fig. 1b). In this geometry the sample showed a giant TMR of ~100%, as shown in Fig. 1b. We also measured microwave power from the MTJ arising from the thermal fluctuations of the free-layer

magnetization^{24,25}. The power was measured by a spectrum analyser, by passing a d.c. current of 1 mA using a bias T .

The radio frequency (r.f.) response of the MTJ was first tested using a network analyser. The results obtained showed evidence of magnetic resonance excited by r.f. current (results not shown).

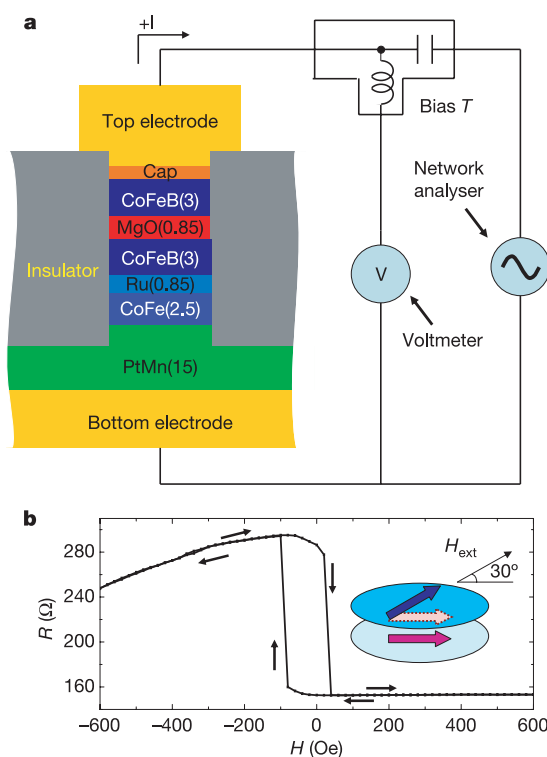


Figure 1 | Experimental set-up and magnetoresistance. **a**, Schematic diagram of the experimental set-up and cross-sectional view of the magnetic tunnel junction (MTJ) device. The thicknesses of various layers of the device in nanometres are given in brackets. The bottom CoFeB and CoFe layers, coupled anti-ferromagnetically through the Ru layer, act as a pinned layer. The top CoFeB layer acts as a free layer, the magnetization of which can be changed. The pinned and free layers are separated by a tunnelling MgO barrier. The experimental set-up measures the d.c. voltage produced across the device on applying the r.f. current. **b**, The magnetoresistance of the device, by applying magnetic field at 30° from the pinned-layer magnetization. The arrows indicate the sweeping direction of the magnetic field.

¹Nanoelectronics Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan. ²CREST, Japan Science and Technology Agency (JST), 4-1-8 Honcho, Kawaguchi 332-0012, Japan. ³Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan.

⁴Electron Device Equipment Division, Canon ANELVA Corporation, 5-8-1 Yotsuya, Fuchu, Tokyo 183-8508, Japan. [†]Present address: Stanford Linear Accelerator Center, Stanford University, California 94025, USA.

Current-induced resonance has recently been observed also in a magnetic domain wall in the megahertz frequency range²⁶. The signal is, however, small, and its phase determination is prone to errors in the calibration of the network analyser. In contrast, we found that the MTJ produces d.c. voltage because of its nonlinear behaviour and this diode-effect measurement offers phase-problem-free results. The effect of alternating current (a.c.) on the precession of magnetization induced by large d.c. current has recently¹⁹ been studied. However, in the present experiment, we excite the magnetization only with alternating current, without applying a d.c. bias. The d.c. voltage response measured by passing 0.55 mA of r.f. current is plotted in Fig. 2. The response shows a large resonance structure, whose position depends on the magnetic field. Figure 3a shows the noise power spectra having a large peak, along with a small side peak, the positions of which also depend on the magnetic field.

The working principles of the spin-torque diode and the semiconductor p–n junction diode are compared in Fig. 4a. As shown in Fig. 4a, when current flows from the n side to the p side, the space charge region around the p–n junction is enlarged, and so the resistance of the p–n junction is higher in this case. For the opposite direction of current, the space charge region is shrunk, which gives lower resistance. In the case of the spin-torque diode, the alternating current passing through it exerts a torque on the free-layer spin moment. When the frequency of the alternating current nears the precession frequency of the free-layer spin-moment, the spin is tilted towards the pinned-layer magnetization during the negative (or positive) half of the alternating current. This configuration has low resistance. During the next half of the alternating current, spin is tilted in the opposite direction, which is a high-resistance state. The difference in resistance during positive and negative currents produces d.c. voltage, in the case of both the diodes. In contrast to the semiconductor diode, the spin-torque diode is resistant to the noise because it produces d.c. voltage only in a narrow frequency range around the resonance frequency, which can be tuned by applying a magnetic field.

Furthermore, the spin-torque diode effect is phase-sensitive. A small r.f. current ($I = I_{\text{a.c.}} \sin(2\pi ft)$) passing through MTJ exerts a torque on the free-layer spin. As a result, the z component (parallel to the pinned-layer magnetization) of the spin oscillates as

$S_z = \cos(\theta) + A \sin(2\pi ft) + B \cos(2\pi ft)$, where A and B are respectively the in-phase and 90°-phase components of the motion of the spin, S_z denotes the direction cosine of the spin and θ is the angle between the free and pinned layers. The resistance of the MTJ depends on S_z as follows: $R = R_0 + 0.5\Delta R(1 - S_z)$, where R_0 is the resistance of the sample when the free layer is parallel to the pinned layer, and ΔR is the increase in the resistance when the free layer is anti-parallel to the pinned layer. The d.c. voltage given by the time-averaged value of $I^2 R$ is $V_{\text{d.c.}} = -A\Delta R I_{\text{a.c.}}^2 / 4$. Thus alternating current passing through the MTJ produces d.c. voltage, which is sensitive to the in-phase part of the oscillation of the magnetic moment. Because this is an intra-sample detection, it gives a phase-error-free spectrum, and a new method for performing ferromagnetic resonance experiments, for which at present an oscillating magnetic field is usually applied.

A torque exerted on the free-layer spin can in general be decomposed into two directions orthogonal to it, that is, $\hat{s}_{\text{free}} \times (\hat{s}_{\text{free}} \times \hat{s}_{\text{pin}})$ and $(\hat{s}_{\text{free}} \times \hat{s}_{\text{pin}})$, where $\hat{s}_{\text{free}}$ and $\hat{s}_{\text{pin}}$ are unit vectors along the magnetizations of the free and pinned layers respectively. The torque from the spin-transfer effect^{1,2} lies in the first direction. The torque along the second direction is called the field-like term, and has been predicted in different ways^{3,4}. Owing to the symmetry differences between these terms (when the free- and pinned-layer magnetizations are in-plane), the d.c. voltage produced by the spin-transfer torque shows a peak, whereas the d.c. voltage produced by an effective field shows a dispersion curve, as shown in Fig. 4b. If both torques act on the spin, by superposition, we get d.c. voltage, as shown in the bottom panel of Fig. 4b. For small and uniform oscillation of the magnetic moment, the d.c. voltage is given approximately by (see Supplementary Discussion):

$$V_{\text{d.c.}} \approx \frac{1}{4} \Delta R \sin^2(\theta) I_{\text{a.c.}}^2 \operatorname{Re} \left[\frac{if\gamma' \text{ST} - \gamma' H_d \gamma' \text{FT}}{(f_0^2 - f^2) - if\alpha\gamma' H_d} \right] \quad (1)$$

where ST is the spin transfer and FT is the effective field term, per unit current. γ is the gyromagnetic ratio ($\gamma' = -\gamma/2\pi > 0$), α is the Gilbert damping factor, H_d is the demagnetization field perpendicular to the free-layer plane, f_0 is the resonant frequency and f is the frequency of the applied alternating current, $I_{\text{a.c.}}$. (We have neglected the stray field effect for a uniform mode.)

The comparison of the d.c. voltage and thermal noise power spectrum at 300 Oe is shown in Fig. 3b. Each spectrum is composed of a large resonance centred at about 7.5 GHz and a small resonance at lower frequency. Here we consider only the dominant resonance (larger peak in the noise) as the uniform oscillation of the free layer (see Supplementary Fig. 1). The d.c. voltage from this mode corresponds to a combination of spin-transfer and effective-field torques (compare with the bottom trace in Fig. 4b). The resonance frequency obtained from fitting to a superposition-type spectrum coincides with the main peak position in the thermal noise spectrum. The frequency of the resonance position as a function of magnetic field is plotted in Fig. 3c. For small θ , and neglecting the marginal influence of d.c. bias current, the resonance frequency is approximately given by Kittel's equation²⁷: $f = \gamma'[(H_c + |H_{\text{dip}} + H_{\text{ext}}|)(H_c + H_d + |H_{\text{dip}} + H_{\text{ext}}|)]^{1/2}$ where H_c is the coercivity and H_{dip} is the dipolar field from the pinned layer. The fitting to this equation gives $H_c + H_{\text{dip}} = 176$ Oe and $H_d = 12.8$ kOe.

As mentioned above, because d.c.-measurement is phase-sensitive detection, we can decompose the spectra into two sources of torque—spin transfer and effective field—by using equation (1); see Supplementary Fig. 2. Further, we also measured the d.c. response by changing the r.f. current (see Fig. 3d), and found that the d.c. voltage decreased linearly with the square of the current, as expected from equation (1).

The d.c. voltage output of an ideal semiconductor diode is given by $V_{\text{d.c.}} \approx eV_{\text{r.f.}}^2/k_B T$, where $V_{\text{r.f.}}$ is the high-frequency voltage applied to the diode, k_B is Boltzmann's constant and T is the temperature. The

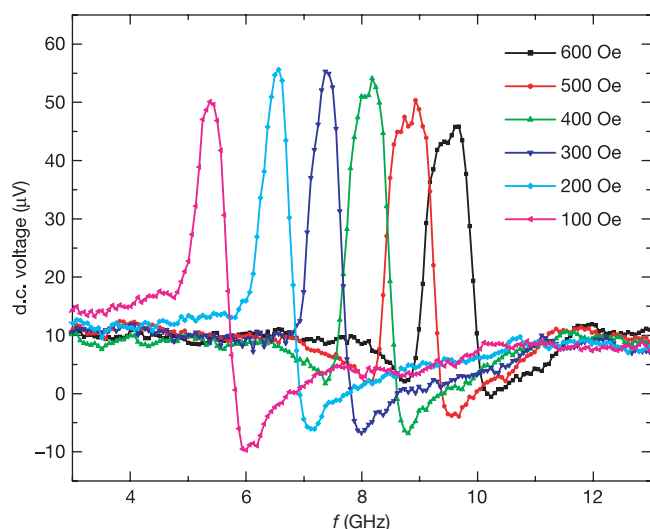


Figure 2 | Direct-current voltage generated by the device in response to the alternating current. The d.c. voltage is plotted as a function of the frequency of the a.c. current (0.55 mA). The external magnetic fields are as shown. The d.c. voltage results from the resonant oscillation of the magnetic moment of the free layer by current-induced spin-transfer and effective-field torques.

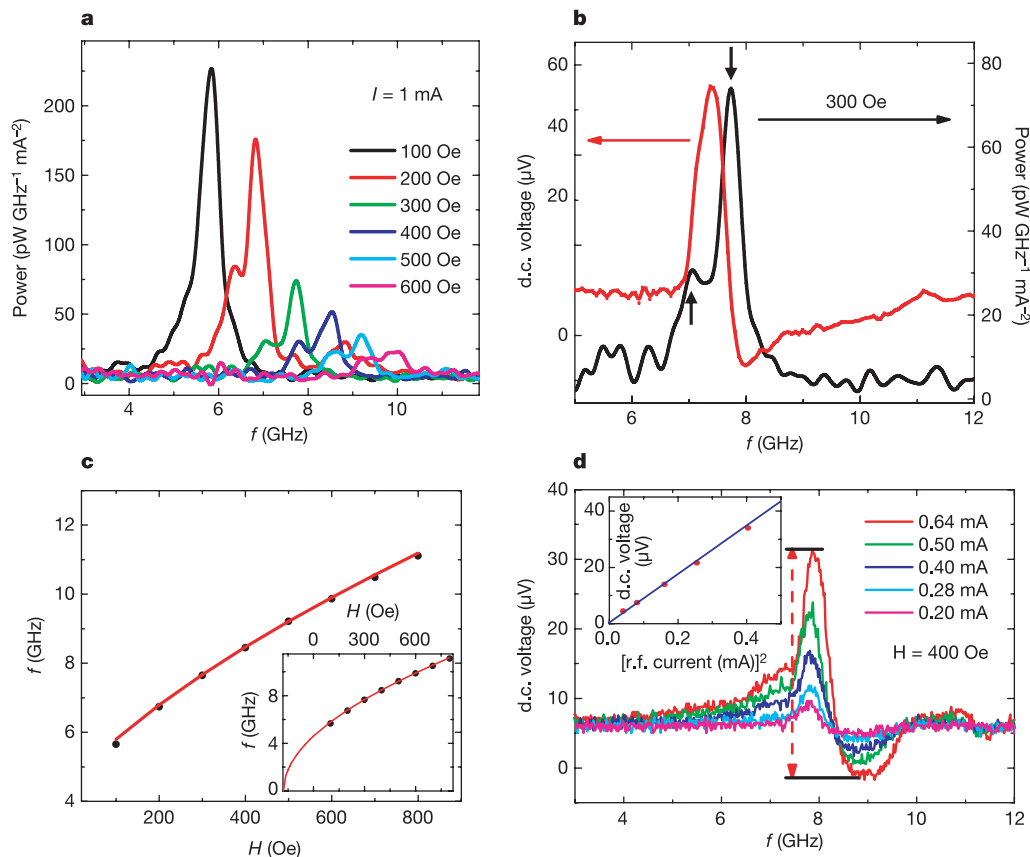


Figure 3 | Magnetic-field dependence of microwave power and a.c. current dependence of d.c. voltage. **a**, Microwave power spectra of the device measured by spectrum analyser, using a d.c. bias current of 1 mA. The power spectra with zero d.c. current are taken as background and subtracted from the data. The external magnetic fields are as shown. **b**, Comparison of the d.c. voltage spectrum and microwave power spectrum at 300 Oe. The arrows mark the positions of two peaks in the microwave power. The d.c. spectrum also has two corresponding resonance frequencies. The shape of the d.c. spectrum is a combination of peak and dispersion curves (also see Fig. 4b).

Thus, the resonance positions do not match with the maxima in the d.c. voltage. **c**, The magnetic-field dependence of resonant frequency corresponding to the larger peak in the microwave power spectra. The black points represent the experimental data points, while the red curve is the fit to the data using Kittel's equation. The inset shows the extrapolation of the fitted curve to zero frequency. **d**, Direct-current voltage from a different sample for the given a.c. currents. The inset shows the linear power dependence of the d.c. voltage measured from peak to valley, as marked by the arrow in the main panel.

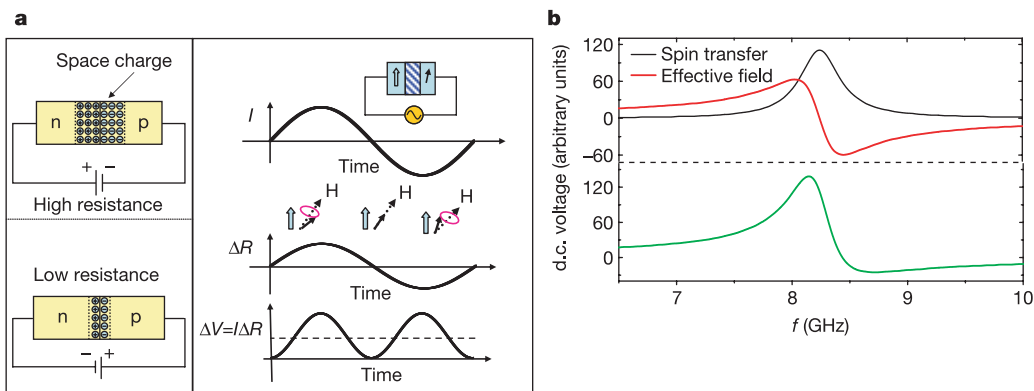


Figure 4 | Principle of the spin-torque diode. **a**, Comparison of the functioning of the semiconductor p-n diode (left panel) and the spin-torque diode (right panel). In the semi-conductor diode, if positive voltage is applied to the n side, the space charge region around the p-n junction is enlarged, and the resistance is high. For the opposite polarity, the space charge region is shrunk and the resistance is low. In the case of the spin-torque diode, the free-layer magnetization (shown by thin black arrows) oscillates owing to the current-induced torque. The resistance of the diode is less when the a.c. current is negative, because the free layer makes a smaller angle with the pinned layer (shown by the thick blue arrow). When the alternating current is positive, the resistance is larger, owing to the larger

angle. The bottom trace in the right panel of **a** shows the schematic variation of the product of the current and the change in resistance. The dotted line shows the average value of this product, which appears as d.c. voltage across the spin-torque diode. **b**, Theoretical plot of the d.c. voltage spectrum obtained from equation (1). The d.c. voltage shows a peak if the torque induced by the a.c. current is due to the spin-transfer effect only. This is shown by the black curve in the top panel. But if the torque is due to the effective-field effect only, the d.c. voltage shows a dispersion curve, as shown by the red curve. If both the torques apply, the d.c. voltage shows a superposition of peak and dispersion, as shown by the green curve in the bottom panel.

spin-torque diode can perform r.f. detection better than the semiconductor diode if the magnetization of the pinned layer lies 45° out-of-plane. In this geometry, by increasing the ratio (H_d/H_c) of free-layer magnetization, we can squeeze the elliptical trajectory of the precession of the magnetic moment to lie in-plane. This enhances the amplitude of in-plane oscillation, and produces large d.c. voltage. The maximum d.c. voltage produced by spin-transfer in this case is given by (see Supplementary Discussion):

$$V_{\text{d.c.}} \approx \text{TMR} \frac{V_{\text{r.f.}}^2}{V_c} \sqrt{\frac{H_d}{H_c}}$$

where V_c is the critical voltage required to flip the magnetization¹. Thus, by increasing the H_d/H_c ratio as well as by increasing the magnetoresistance (as we have done here using a high-quality crystalline MgO barrier), the spin-torque diode can be a sensitive power detector. The sensitivity can also be enhanced by increasing the spin-transfer efficiency^{28,29}, which decreases the critical voltage.

Received 28 March; accepted 1 September 2005.

- Slonczewski, J. C. Current-driven excitation of magnetic multilayers. *J. Magn. Magn. Mater.* **159**, L1–L7 (1996).
- Berger, L. Emission of spin waves by a magnetic multilayer traversed by a current. *Phys. Rev. B* **54**, 9353–9358 (1996).
- Zhang, S., Levy, P. M. & Fert, A. Mechanisms of spin-polarized current-driven magnetization switching. *Phys. Rev. Lett.* **88**, 236601 (2002).
- Tatara, G. & Kohno, H. Theory of current-driven domain wall motion: spin transfer versus momentum transfer. *Phys. Rev. Lett.* **92**, 086601 (2004).
- Kiselev, S. I. *et al.* Microwave oscillations of a nanomagnet driven by a spin-polarized current. *Nature* **425**, 380–383 (2003).
- Rippard, W. H., Pufall, M. R., Kaka, S., Russek, S. E. & Silva, T. J. Direct-current induced dynamics in $\text{Co}_{90}\text{Fe}_{10}/\text{Ni}_{80}\text{Fe}_{20}$ point contacts. *Phys. Rev. Lett.* **92**, 027201 (2004).
- Sun, J. Applied physics: spintronics gets a magnetic flute. *Nature* **425**, 359–361 (2003).
- Myers, E. B., Ralph, D. C., Katine, J. A., Louie, R. N. & Buhrman, R. A. Current-induced switching of domains in magnetic multilayer devices. *Science* **285**, 867–870 (1999).
- Tsoi, M. *et al.* Generation and detection of phase-coherent current-driven magnons in magnetic multilayers. *Nature* **406**, 46–48 (2000).
- Tulapurkar, A. A. *et al.* Subnanosecond magnetization reversal in magnetic nanopillars by spin angular momentum transfer. *Appl. Phys. Lett.* **85**, 5358–5360 (2004).
- Yagami, K., Tulapurkar, A. A., Fukushima, A. & Suzuki, Y. Low-current spin-transfer switching and its thermal durability in a low-saturation-magnetization nanomagnet. *Appl. Phys. Lett.* **85**, 5634–5636 (2004).
- Koch, R. H., Katine, J. A. & Sun, J. Z. Time-resolved reversal of spin-transfer switching in a nanomagnet. *Phys. Rev. Lett.* **92**, 088302 (2004).
- Urazhdin, S., Birge, N. O., Pratt, W. P. Jr & Bass, J. Current-driven magnetic excitations in permalloy-based multilayer nanopillars. *Phys. Rev. Lett.* **91**, 146803 (2003).
- Özyilmaz, B. *et al.* Current-induced magnetization reversal in high magnetic fields in Co/Cu/Co nanopillars. *Phys. Rev. Lett.* **91**, 067203 (2003).
- Grollier, J. *et al.* Spin-polarized current induced switching in Co/Cu/Co pillars. *Appl. Phys. Lett.* **78**, 3663–3665 (2001).
- Katine, J. A., Albert, F. J., Buhrman, R. A., Myers, E. B. & Ralph, D. C. Current-driven magnetization reversal and spin-wave excitations in Co/Cu/Co pillars. *Phys. Rev. Lett.* **84**, 4212–4215 (2000).
- Stiles, M. D. & Zangwill, A. Anatomy of spin-transfer torque. *Phys. Rev. B* **66**, 014407 (2002).
- Lee, K.-J., Deac, A., Redon, O., Nozières, J.-P. & Dieny, B. Excitations of incoherent spin-waves due to spin-transfer torque. *Nature Mater.* **3**, 877–881 (2004).
- Pufall, M. R., Rippard, W. H., Kaka, S., Silva, T. J. & Russek, S. E. Frequency modulation of spin-transfer oscillators. *Appl. Phys. Lett.* **86**, 082506 (2005).
- Yuasa, S., Nagahama, T., Fukushima, A., Suzuki, Y. & Ando, K. Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions. *Nature Mater.* **3**, 868–871 (2004).
- Parkin, S. S. P. *et al.* Giant tunnelling magnetoresistance at room temperature with MgO (100) tunnel barriers. *Nature Mater.* **3**, 862–867 (2004).
- Djayaprawira, D. D. *et al.* 230% Room temperature magnetoresistance in CoFeB/MgO/CoFeB magnetic tunnel junctions. *Appl. Phys. Lett.* **86**, 092502 (2005).
- Sze, S. M. *Physics of Semiconductor Devices* Ch. 2 (John Wiley & Sons, New York, 1981).
- Smith, N. & Arnett, P. White-noise magnetization fluctuations in magnetoresistive heads. *Appl. Phys. Lett.* **78**, 1448–1450 (2001).
- Nazarov, A. V., Cho, H. S., Nowak, J., Stokes, S. & Tabat, N. Tunable ferromagnetic resonance peak in tunneling magnetoresistive sensor structures. *Appl. Phys. Lett.* **81**, 4559–4561 (2002).
- Saitoh, E., Miyajima, H., Yamaoka, T. & Tatara, G. Current-induced resonance and mass determination of a single magnetic domain wall. *Nature* **432**, 203–206 (2004).
- Kittel, C. *Introduction to Solid State Physics* Ch. 16 (John Wiley & Sons, Singapore, 1996).
- Jiang, Y. *et al.* Substantial reduction of critical current for magnetization switching in an exchange-biased spin valve. *Nature Mater.* **3**, 361–364 (2004).
- Manschoot, J., Brataas, A. & Bauer, G. E. W. Reducing the critical switching current in nanoscale spin valves. *Appl. Phys. Lett.* **85**, 3250–3252 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements A.A.T. thanks the Japan Society for the Promotion of Science for the fellowship grant. S.Y. thanks the Japan Science and Technology Agency (JST) for the PRESTO programme. A part of this work is supported by the 21st Century COE programme by JSPS. We thank C. Chappert, T. Devolder, W. Mizutani and M. Mizuguchi for their help.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.S. (suzuki-y@mp.es.osaka-u.ac.jp).

Chaos-based communications at high bit rates using commercial fibre-optic links

Apostolos Argyris¹, Dimitris Syvridis¹, Laurent Larger², Valerio Annovazzi-Lodi³, Pere Colet⁴, Ingo Fischer^{5†}, Jordi García-Ojalvo⁶, Claudio R. Mirasso⁷, Luis Pesquera⁸ & K. Alan Shore⁹

Chaotic signals have been proposed as broadband information carriers with the potential of providing a high level of robustness and privacy in data transmission^{1,2}. Laboratory demonstrations of chaos-based optical communications have already shown the potential of this technology^{3–5}, but a field experiment using commercial optical networks has not been undertaken so far. Here we demonstrate high-speed long-distance communication based on chaos synchronization over a commercial fibre-optic channel. An optical carrier wave generated by a chaotic laser is used to encode a message for transmission over 120 km of optical fibre in the metropolitan area network of Athens, Greece. The message is decoded using an appropriate second laser which, by synchronizing with the chaotic carrier, allows for the separation of the carrier and the message. Transmission rates in the gigabit per second range are achieved, with corresponding bit-error rates below 10^{-7} . The system uses matched pairs of semiconductor lasers as chaotic emitters and receivers, and off-the-shelf fibre-optic telecommunication components. Our results show that information can be transmitted at high bit rates using deterministic chaos in a manner that is robust to perturbations and channel disturbances unavoidable under real-world conditions.

Broadband information carriers enhance the robustness of communication channels to interferences with narrow-band disturbances. This is the basis of spread-spectrum communication techniques, such as the code division multiple access (CDMA) protocol used in the Global Positioning System (GPS) and in the third generation of mobile telephones. In chaos-based communications the broadband coding signal is generated at the physical layer instead of algorithmically. Additionally, chaotic carriers offer a certain degree of intrinsic privacy in the data transmission, which could complement (via robust hardware encryption) both classical (software-based)⁶ and quantum⁷ cryptography systems. From a fundamental viewpoint, using waveforms generated by deterministic chaotic systems to carry information in a robust manner that also allows high bit rates is a generalization of standard communication systems. It might also provide a deeper insight into the mechanism of transmission of information in natural systems with complex dynamics, such as biological systems.

Chaotic communication systems based on chaos synchronization⁸ were proposed in the early 1990s (refs 1, 2). In this type of communication protocol, messages are embedded within a chaotic carrier in the emitter, and recovered after transmission by a receiver upon synchronization with the emitter. The receiver architecture can be viewed as performing a nonlinear filtering process, intended to

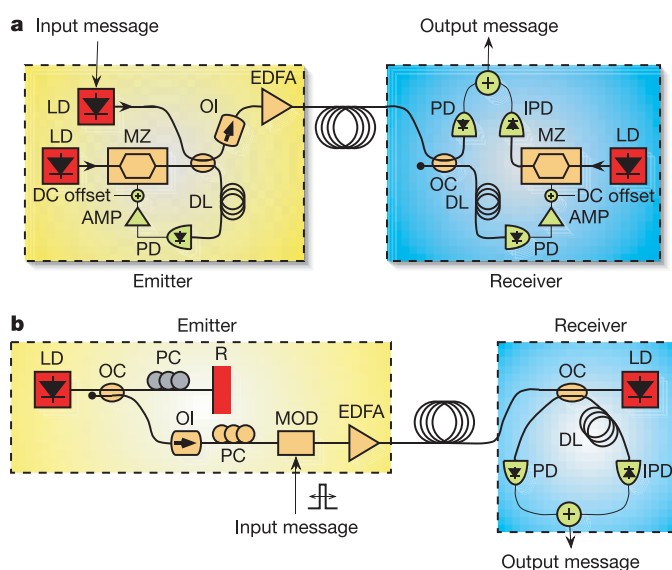


Figure 1 | Two schematic set-ups for optical chaos communication. In the optoelectronic scheme (a), the emitter is a laser diode (LD) whose output is modulated in a strongly nonlinear way by an electro-optic feedback loop through an integrated electro-optic Mach-Zehnder interferometer (MZ). The message is added inside the delay oscillation loop. An erbium-doped fibre amplifier (EDFA) is used to adjust the power to be injected into the transmission line. The EDFA is followed by an optical filter (not shown) that cuts off the amplified spontaneous emission noise. All LDs operate at around $1.55\ \mu\text{m}$. In the all-optical scheme (b), the emitter is a laser diode subject to optical feedback from a digital variable reflector (R). The length of the external cavity is 6 m; a polarization controller (PC) is used within the cavity to adjust the polarization state of the light reflected back from the variable reflector. The message is added via a modulator (MOD) at the emitter's output. A typical transmission module, represented by a fibre loop in the figure, consists of a combination of single-mode and dispersion-compensated fibres, followed by an EDFA that compensates for the power lost upon transmission. In both schemes, decoding is performed via subtraction of the transmitted signal from the signal filtered by the receiver. Operationally, the subtraction is performed by adding the photocurrents coming from an ordinary and a sign-inverting amplified photodiode (PD and IPD, respectively). Fibre connections are represented by thick lines, and electric connections by thin lines. Other elements in the diagram include optical isolators (OI), delay lines (DL), electronic amplifiers (AMP) and optical fibre couplers (OC).

¹Department of Informatics, University of Athens, Athens 15784, Greece. ²UMR CNRS 6174, FEMTO-ST / Optics, University of Franche-Comté, 16 route de Gray, 25030 Besançon cedex, France. ³Department of Electronics, University of Pavia, 27100 Pavia, Italy. ⁴Instituto Mediterráneo de Estudios Avanzados, (IMEDEA, CSIC-UIB), Campus UIB, E-07122 Palma de Mallorca, Spain. ⁵Institute of Applied Physics, Darmstadt University of Technology, 64289 Darmstadt, Germany. ⁶Departament de Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, E-08222 Terrassa, Spain. ⁷Department de Física, Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain. ⁸Instituto de Física de Cantabria (CSIC-UC), Avd de los Castros s/n, E-39005 Santander, Spain. ⁹School of Informatics, University of Wales, Bangor LL57 1UT, UK. †Present address: Department of Applied Physics and Photonics, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.

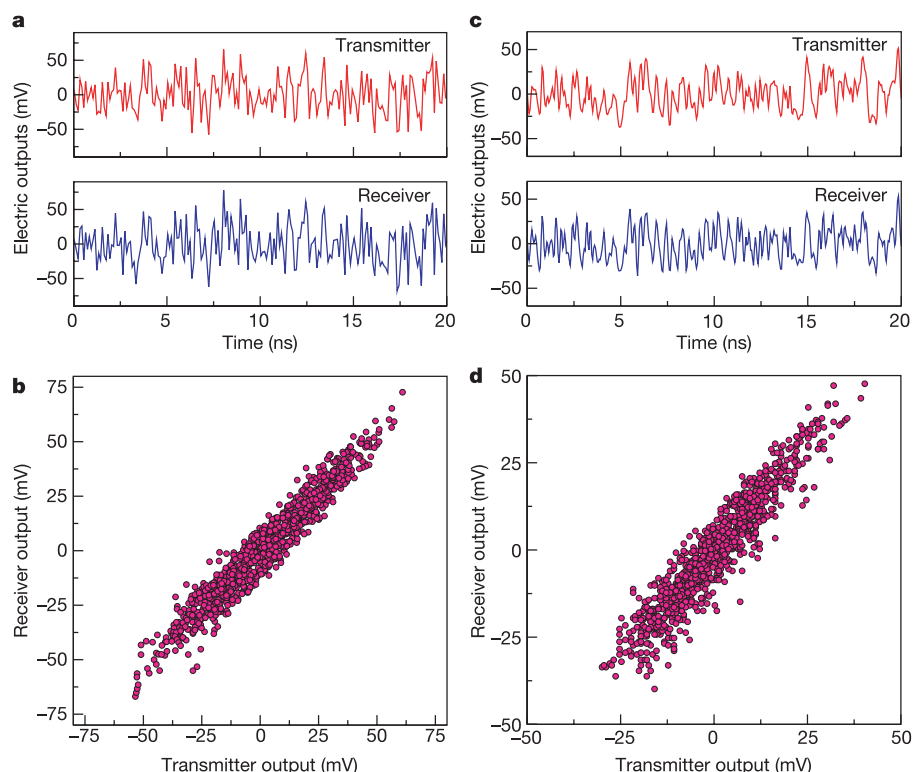


Figure 2 | Back-to-back synchronization. Time series (**a** and **c**) and synchronization plots, receiver output power versus transmitter output power (**b** and **d**), of two unidirectionally back-to-back coupled electro-

optical (**a**, **b**) and all-optical (**c**, **d**) systems. Configurations correspond to those depicted in Fig. 1. **b**, **d**, Plots contain 1,000 points, sampled every 100 and 50 ps, respectively.

generate locally a message-free chaotic signal, which is then used for subtraction from the encoded transmitted signal.

Optical systems provide simple ways of generating very high-dimensional chaotic carriers that offer a substantial security level, and also the possibility of very high transmission rates⁹. Early laboratory experiments demonstrated successful back-to-back communications in all-optical³ and electro-optical⁴ systems. High bit rates have been achieved in these back-to-back conditions^{5,10}, but no long-distance transmission experiments in commercial communication networks have been undertaken so far. Here we report the achievement of chaotic optical communications over long fibre spans (more than 100 km), at high transmission rates (higher than 1 Gb s^{-1}) and low bit-error rates (lower than 10^{-7}). We report results from a successful field experiment using part of the metropolitan area network of the city of Athens, Greece.

Generation of chaotic signals with high dimension and high information entropy can be achieved in diode lasers by means of delayed feedback. We have considered two schemes, involving all-optical and electro-optical feedback. In the first scheme (Fig. 1a), the output of a diode laser is used to drive an integrated Mach-Zehnder interferometer that acts as a nonlinear intensity modulator; in the second (Fig. 1b), a diode laser is subject to coherent optical feedback from an external mirror. In both set-ups, a high-dimensional chaotic output is obtained in a large region of parameter space^{11,12}. Distributed-feedback lasers are used to ensure single-mode operation.

In the electro-optical scheme encoding is realized by adding the message to the chaotic carrier inside the nonlinear feedback loop (Fig. 1a), whereas in the all optical scheme the chaotic carrier amplitude is modulated with the message (Fig. 1b). In both schemes, the message is decoded by subtracting the total transmitted signal from the output of the receiver laser, which is not subject to feedback (hence the receiver is not chaotic in the absence of the transmitted signal). This architecture is known as an 'open-loop' configuration,

as opposed to the 'closed-loop' scheme in which both transmitter and receiver have feedback. Earlier studies have shown that the open-loop configuration leads to synchronization more readily than the closed-loop one^{13,14}. Furthermore, in the all-optical case the closed-loop architecture requires careful tuning of the feedback phase^{15,16}.

First we examine the back-to-back performance of the electro-optical communication set-up (see Fig. 1a). The emitter and receiver in this case are two nominally identical distributed-feedback

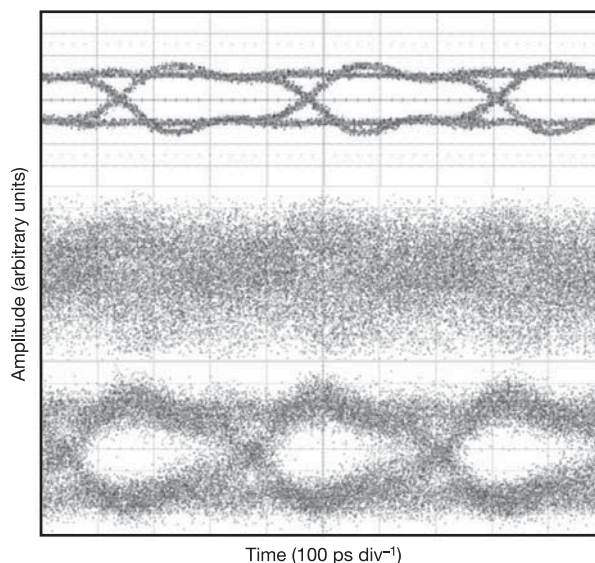


Figure 3 | Representative eye diagrams in the electro-optic set-up. The message is shown in the top trace, the encoded signal in the middle trace, and the decoded message in the bottom trace.

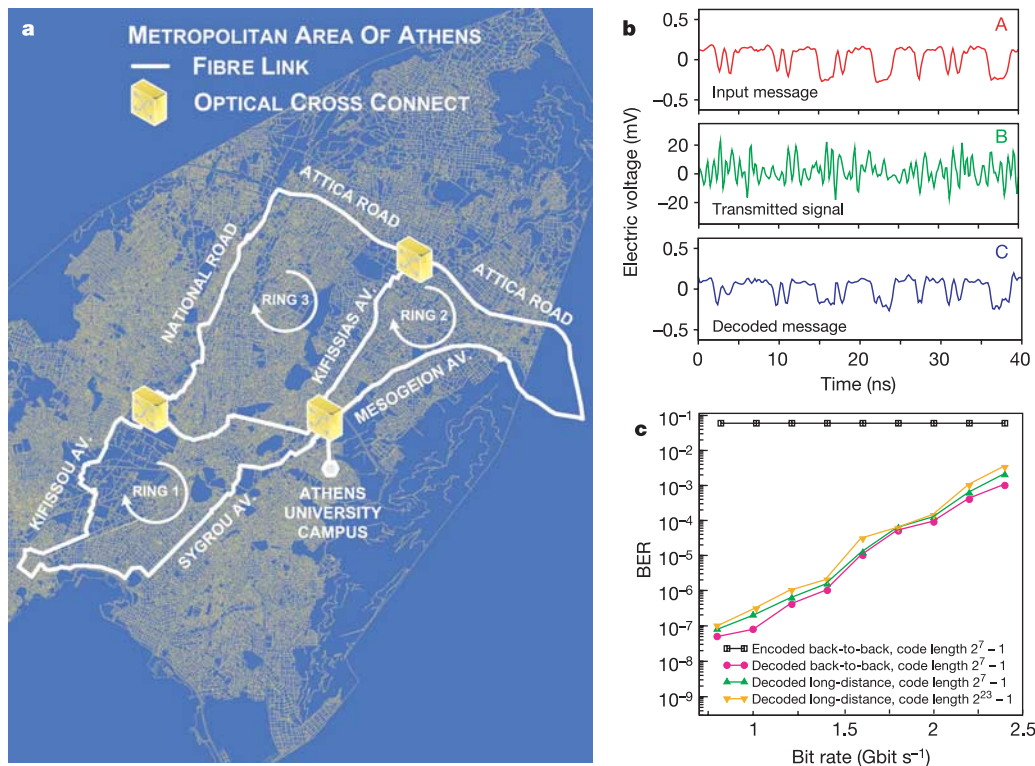


Figure 4 | Field experiment of fibre transmission. **a**, Implementation of chaos-encoded communications in the optical communication network of Athens, Greece. **b**, Time traces of a 1 Gb s^{-1} applied message (trace A; $\text{BER} < 10^{-12}$), carrier with the encoded message (trace B; $\text{BER} \approx 6 \times 10^{-2}$)

and recovered message after 120-km transmission (trace C; $\text{BER} \approx 10^{-7}$). **c**, The BER performance of the encoded signal (squares), back-to-back decoded message (circles) and decoded message after transmission for two different code lengths (triangles).

semiconductor lasers operating at 1552.0 and 1552.9 nm, respectively. The electro-optical feedback produces a chaotic carrier signal with a broad power spectrum. Upon injection of the emitter output into the receiver (which does not have feedback of its own), the receiver also becomes chaotic and synchronizes with the former, as shown in Fig. 2a, b. The corresponding behaviour of the back-to-back all-optical set-up (see Fig. 1b) is represented in Fig. 2c, d. In this case the emission wavelengths of the two lasers have been matched with proper temperature adjustment. A clear synchronization between the emitter and receiver outputs is observed in this case as well.

Synchronization is preserved after propagation through the fibre (results not shown). In particular, we have performed laboratory experiments with two transmission modules consisting of 50 km of single-mode fibre and 6 km of dispersion compensation fibre each. The compensation module was selected to have a total dispersion around -850 ps nm^{-1} , to counterbalance the dispersion of the single-mode fibre (around 850 ps nm^{-1}). Each module is completed with an erbium-doped fibre amplifier whose gain is tuned to provide the necessary power recovery after transmission, followed by an optical filter to remove unwanted amplified spontaneous emission noise.

Figure 3 shows the performance of the electro-optical set-up after transmission. A message in the form of a pseudorandom bit sequence of $2^7 - 1$ bits is introduced via chaos modulation, as explained above (see Fig. 1a). The bias voltage of the modulator is chosen so that either the 'ones' or 'zeros' have a flattened noise profile, to avoid low-frequency noise that degrades the bit-error rate. Figure 3 shows the eye diagrams (superposition of a large number of bits) of the message and of the encoded and decoded signals after propagation through the transmission modules described in the previous paragraph. Bit-error rates (BER) are of the order of 10^{-7} , for a transmission rate of 3 Gb s^{-1} . The performance of the all-optical

scheme in laboratory experiments is similar in terms of BER (see below), for slightly smaller transmission rates (of the order of 1 Gb s^{-1}).

To test performance under 'real-world' conditions we implemented a chaos-based all-optical transmission system using an installed optical network infrastructure of single-mode fibre belonging to the metropolitan area network of Athens. The network has a total length of 120 km and was provided by Attika Telecom SA. The topology consists of three fibre rings, linked together at specific cross-connect points (see Fig. 4a). Through three cross-connect points, the transmission path follows the Ring-1 route, then the Ring-2 route, and finally the Ring-3 route. A dispersion compensation fibre module, set at the beginning of the link (pre-compensation technique), cancels the chromatic dispersion that would be induced by the single-mode fibre transmission. Erbium-doped fibre amplifiers and optical filters are used along the optical link for compensation of the optical losses and filtering of amplified spontaneous emission noise, respectively. The pair of lasers is selected to exhibit parameter mismatches that are constrained below 3%. The mean optical power injected into the receiver has been limited to 0.8 mW, to avoid possible damage of the antireflective coating of the slave laser.

A non-return-to-zero pseudorandom bit sequence is applied by externally modulating the chaotic carrier by means of a LiNbO_3 Mach-Zehnder modulator. The message amplitude is attenuated 14 dB with respect to the carrier, so the BER of the transmitted signal after filtering (but without the appropriate decoding) is always larger than 6×10^{-2} , this value being the instrumentation limit (the maximum BER of the error detector is 9×10^{-2}). A good synchronization performance of the transmitter-receiver set-up leads to an efficient cancellation of the chaotic carrier, and hence to a satisfactory decoding process (see Fig. 4b). The performance of the chaotic transmission system has been studied for different message bit rates up to 2.4 Gb s^{-1} and for two different code lengths: $2^7 - 1$

and $2^{23} - 1$. All BER values have been measured after filtering the subtracted electric signal, by using low-pass filters with bandwidth adjusted each time to the message bit rate. For transmission rates in the gigabit per second range the recovered message exhibits BER values lower than 10^{-7} . For higher transmission rates the corresponding bit-error rates increase, as shown in Fig. 4c. This is due to the fact that the synchronization is not perfect, mainly because of parameter differences between the two lasers. Moreover, signal extraction is less efficient at bit rates comparable to the relaxation oscillation frequencies of the lasers (in our case around 3 GHz). For properly chosen lasers (that is, with much higher relaxation frequency and better parameter matching) this BER deterioration could be avoided to a large extent, and BER values could approach the ones of traditional communication systems, which lie between 10^{-9} and 10^{-12} .

When the code length increases from $2^7 - 1$ to $2^{23} - 1$ bits, only a minor increase in BER is observed. Also, relatively small differences in the BER values exist between the back-to-back and the transmission set-ups, revealing only a slight degradation of the system performance due to the transmission link. Communication bit rates are mainly limited by the bandwidth of the chaotic carrier, which depends on the emitter components. Nevertheless, this bandwidth can extend well beyond the relaxation frequency of the emitter laser. In our electro-optical set-up the bandwidth is about 7 GHz, and in the all-optical set-up it is 5 GHz.

The results reported here provide a convincing proof-of-practical-concept for optical chaos communications technology. Building on this, it should be possible to develop reliable cost-effective secure communications systems that exploit deeper properties of chaotic dynamics¹⁷. Opportunities for such technological advances have emerged from the substantial theoretical and experimental advances accomplished within the OCCULT project (<http://nova.uib.es/project/occult>) and in other laboratories around the world^{18–20}.

The prototype technology that has been developed would benefit from added intelligence features. It will be essential to develop a meaningful measure of the security provided by any proposed secure communications technology—one capable of useful comparison with other technologies. Beyond that, we envisage opportunities for effecting fundamental advances in chaos synchronization techniques, using smart encryption techniques¹⁶, developing active eavesdropper-evasion strategies, designing compact transmitter/receiver modules, as well as implementing robust technology for bidirectional chaotic communications, chaotic message broadcasting²¹, parallel communications with spatiotemporal chaos^{22,23}, and frequency multiplexing through sharing of the chaotic broadband spectrum²⁴. In addition, chaos communication systems are fully compatible with the widely used wavelength division multiplexing (WDM), provided the chaotic carrier bandwidth is much smaller than the WDM channel spacing. In combination, such advances should enable the delivery of practical systems for intelligent chaotic optical data encoding and lead to deeper fundamental insight into communication between systems with irregular or even adaptive signals.

Received 11 July; accepted 26 September 2005.

1. Cuomo, K. M., Oppenheim, A. V. & Strogatz, S. H. Synchronization of Lorenz-based chaotic circuits with applications to communications. *IEEE Trans. Circuits Syst. II* **40**, 626–633 (1993).

2. Colet, P. & Roy, R. Digital communication with synchronized chaotic lasers. *Opt. Lett.* **19**, 2056–2058 (1994).
3. Van Wiggeren, G. D. & Roy, R. Communications with chaotic lasers. *Science* **279**, 1198–1200 (1998).
4. Goedgebuer, J. P., Larger, L. & Porte, H. Optical cryptosystem based on synchronization of hyperchaos generated by a delayed feedback tunable laser diode. *Phys. Rev. Lett.* **80**, 2249–2252 (1998).
5. Tang, S. & Liu, J. M. Message encoding-decoding at 2.5 Gbits/s through synchronization of chaotic pulsing semiconductor lasers. *Opt. Lett.* **26**, 1843–1845 (2001).
6. Shannon, C. E. Communication theory of secrecy systems. *Bell Syst. Technic. J.* **28–4**, 656–715 (1949).
7. Kertsiefer, C. et al. A step towards global key distribution. *Nature* **419**, 450 (2002).
8. Ashwin, P. Synchronization from chaos. *Nature* **422**, 384–385 (2003).
9. Mirasso, C. R., Colet, P. & Garcia-Fernandez, P. Synchronization of chaotic semiconductor lasers: Application to encoded communications. *IEEE Photon. Technol. Lett.* **8**, 299–301 (1996).
10. Kusumoto, K. & Ohtsubo, J. 1.5-GHz message transmission based on synchronization of chaos in semiconductor lasers. *Opt. Lett.* **27**, 989–991 (2002).
11. Annovazzi-Lodi, V., Merlo, S., Norgia, M. & Scirè, A. Characterization of a chaotic telecommunication laser for different fiber cavity lengths. *IEEE J. Quant. Electron.* **38**, 1171–1177 (2002).
12. Lee, M. W., Larger, L. & Goedgebuer, J. P. Transmission system using chaotic delays between lightwaves. *IEEE J. Quant. Electron.* **39**, 931–936 (2003).
13. Vicente, R., Pérez, T. & Mirasso, C. R. Open- versus closed-loop performance of synchronized chaotic external-cavity semiconductor lasers. *IEEE J. Quant. Electron.* **38**, 1197–1204 (2002).
14. Lee, M. W., Paul, J., Sivaprakasam, S. & Shore, K. A. Comparison of closed-loop and open-loop feedback schemes of message decoding using chaotic laser diodes. *Opt. Lett.* **28**, 2168–2170 (2003).
15. Peil, M., Heil, T., Fischer, I. & Elsässer, W. Synchronization of chaotic semiconductor laser systems: a vectorial coupling-dependent scenario. *Phys. Rev. Lett.* **88**, 174101 (2002).
16. Heil, T. et al. ON/OFF phase shift keying for chaos-encrypted communication using external-cavity semiconductor lasers. *IEEE J. Quant. Electron.* **38**, 1162–1170 (2002).
17. Annovazzi Lodi, V., Benedetti, M., Merlo, S. & Provinzano, M. N. B. Optical chaos masking of video signals. *IEEE Photon. Technol. Lett.* **17**, 1995–1997 (2005).
18. Donati, S. & Mirasso, C. R. (eds) Feature section on optical chaos and applications to cryptography. *IEEE J. Quant. Electron.* **39**, 1138–1204 (2002).
19. Larger, L. & Goedgebuer, J. P. (eds) Cryptography using optical chaos. *C.R. Phys.* **5**, 609–681 (2004).
20. Uchida, A., Rogister, F., Garcia-Ojalvo, J. & Roy, R. Synchronization and communication with chaotic optical systems. *Prog. Opt.* **48**, 203–341 (2005).
21. Lee, M. W. & Shore, K. A. Chaotic message broadcasting using DFB laser diodes. *Electron. Lett.* **40**, 614–615 (2004).
22. Gollub, J. P. & Cross, M. C. Chaos in space and time. *Nature* **404**, 710–711 (2000).
23. Garcia-Ojalvo, J. & Roy, R. Spatiotemporal communication with synchronized optical chaos. *Phys. Rev. Lett.* **86**, 5204–5207 (2001).
24. Matsuura, T., Uchida, A. & Yoshimori, S. Chaotic wavelength division multiplexing for optical communications. *Opt. Lett.* **29**, 2731–2733 (2004).

Acknowledgements Financial support was provided by the European Commission through the IST project OCCULT. We thank ATTICA TELECOMMS SA for providing the installed fibre infrastructure for the field experiment, and S. Hansmann and J. Schumacher for providing the matched laser and detector pairs. We thank S. H. Strogatz for a careful reading of the manuscript. We are also grateful for the collaboration of J. P. Goedgebuer, W. Elsässer, M. Peil, S. Poinot, M. Benedetti, S. Merlo, M. Norgia, M. W. Lee, Y. Chembo Kouomou, T. Pérez, R. Vicente, J. Mulet, J. M. Buldú, M. C. Torrent, S. Ortin and the rest of the participants of the OCCULT project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.R.M. (claudio@galiota.uib.es).

Global pattern of trends in streamflow and water availability in a changing climate

P. C. D. Milly¹, K. A. Dunne¹ & A. V. Vecchia²

Water availability on the continents is important for human health^{1,2}, economic activity³, ecosystem function⁴ and geophysical processes⁵. Because the saturation vapour pressure of water in air is highly sensitive to temperature, perturbations in the global water cycle are expected to accompany climate warming⁶. Regional patterns of warming-induced changes in surface hydroclimate are complex and less certain than those in temperature, however, with both regional increases and decreases expected in precipitation and runoff. Here we show that an ensemble of 12 climate models exhibits qualitative and statistically significant skill in simulating observed regional patterns of twentieth-century multidecadal changes in streamflow. These models project 10–40% increases in runoff in eastern equatorial Africa, the La Plata basin and high-latitude North America and Eurasia, and 10–30% decreases in runoff in southern Africa, southern Europe, the Middle East and mid-latitude western North America by the year 2050. Such changes in sustainable water availability would have considerable regional-scale consequences for economies as well as ecosystems.

Streamflow is a temporally lagged, spatial integral of runoff over a river basin. Averaged over many years, runoff generally is equal to the difference between precipitation and evapotranspiration and, hence, to the convergence of horizontal atmospheric water flux. From a resource perspective, runoff is a measure of sustainable water availability. However, streamflow can be affected by anthropogenic disturbances, which may generate spurious (that is, nonclimatic) changes; at the spatial scale of basins to be considered here, the most significant of these disturbances is associated with the diversion of water for the irrigation of cropland.

In support of an assessment of forced climate change conducted by the Intergovernmental Panel on Climate Change (IPCC), many climate-modelling centres recently performed '20C3M' simulations of climate with prescribed external forcing (variations in atmospheric composition and solar irradiance) for the late nineteenth century and the whole of the twentieth century. Forcing was not identical across models, but generally included estimated historical variations of radiatively active atmospheric gases and aerosols (including volcanic emissions) and solar irradiance. Control simulations with temporally invariant preindustrial forcing ('PICNTRL') were also performed, as were integrations into the future with an assumed forcing model ('SRESA1B').

The annual runoff fields from a total of 62 runs of the 20C3M experiment on 21 different models (one to nine runs per model) were integrated spatially over 165 river basins with long-term (28–99 years, median 59 years) streamflow measurements (see Methods). Climate-model runoff commonly does not reflect the time lag associated with storage in river basins. Although the missing time lag may not affect computed climates, it can strongly affect the temporal variability of streamflow, which is important for our

analysis. Accordingly, the climate-model basin runoff time series were converted to equivalent streamflow time series by routing them through a model of a linear reservoir in such a way as to match the observed serial correlation (see Methods).

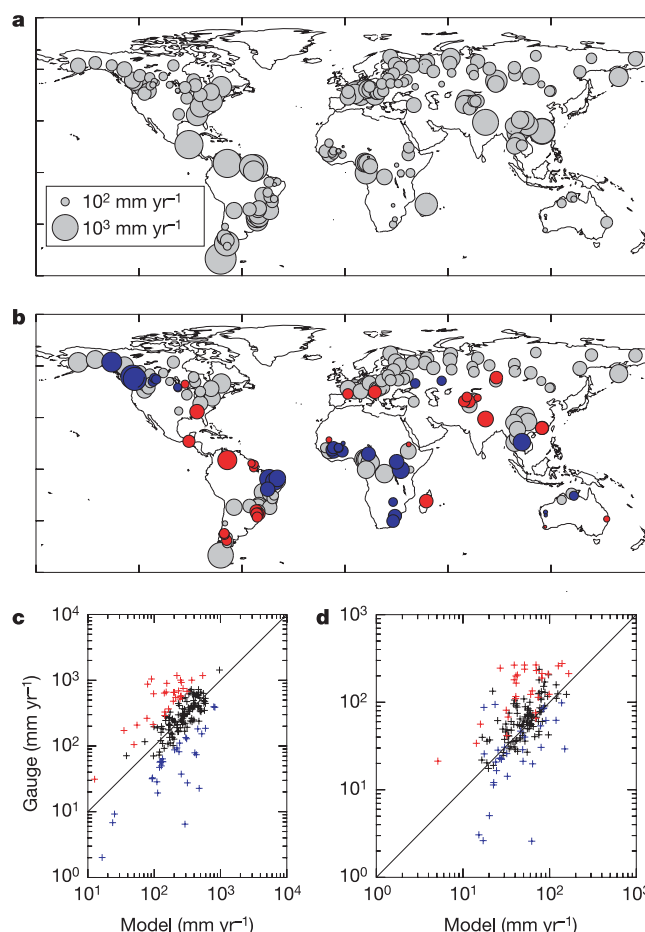


Figure 1 | Annual runoff rate (streamflow per unit basin area, in mm yr^{-1}). **a**, Global distribution of mean values from stream-gauge observations; runoff is proportional to the area of each circle, and each circle is centred at a gauged basin centroid. **b**, Global distribution of ensemble (geometric) means from 35 model runs. Here and in **c** and **d**, gauges with a mean greater than double (smaller than half) the observations are shown in blue (red). (Geometric mean was used because runoff estimates varied greatly across models for a given gauge; results for arithmetic means are similar.) **c**, Observed versus model ensemble means. **d**, Observed versus model ensemble standard deviations.

¹US Geological Survey, Geophysical Fluid Dynamics Laboratory/NOAA, PO Box 308, Princeton, New Jersey 08542, USA. ²US Geological Survey, 821 E. Interstate Ave., Bismarck, North Dakota 58503-1199, USA.

To maximize the signal-to-noise ratio in our analysis, we chose to work with an ensemble of models. The realism of hydroclimatic simulations varies across models, so we elected to form the ensemble from a subset of the models, with the selection based on performance. We ranked the models with respect to root-mean-square (r.m.s.) error (over the 165 basins and all runs) of the logarithm of long-term mean discharge per unit area; the logarithmic transform is commonly used in hydrology because flows can range over orders of magnitude. We retained the 12 models (35 runs of 20C3M) with the lowest error for use in the ensemble analyses presented here (see Methods).

The mean and standard deviation of annual streamflows (expressed per unit basin area) for the ensemble model output generally range from about one-half to double the observed values, except in a few regions (Fig. 1). In relation to the observed values, the ensemble model mean values tend to be large in much of Africa, the Nordeste region of South America (northeastern Brazil), and north-west North America, and small in northern low latitudes of the Americas and southern South America. Despite the presence of these large local-scale differences between the model ensemble and observed values, the global-scale relation is strong.

To characterize twentieth-century changes in streamflow at gauges, we used the difference, D , between the average annual streamflows for 1971–98 (based on available sample size m) and

1900–70 (sample size n) (see Methods). Let s and r be the sample standard deviation and lag-1 autocorrelation of the pooled annual streamflow time series. Under the null hypothesis of a stationary climate, the normalized difference, defined as $Z = D / \{s(1/m + 1/n)^{1/2}[(1+r)/(1-r)]^{1/2}\}$ and henceforth termed the ‘trend’, is approximately standard-normally distributed⁷. Thus, normalizing D in this way accounts for local differences in record lengths and streamflow variability and persistence. Although Z can be used to measure local significance of change, our objective is not to assess local significance but rather to determine whether the observed Z values for the 165 basins are correlated significantly with the values predicted by the 12-model ensemble. If so, we can conclude that external forcing explains a significant part of global streamflow change for the twentieth century.

The pattern of hydroclimatic change indicated by the ensemble mean of the model trends qualitatively resembles the pattern in the observations (Fig. 2). The observed tendency towards less runoff in sub-Saharan Africa, southern Europe, southernmost South America, southern Australia and western mid-latitude North America generally is seen in the ensemble. The ensemble reproduces the observed increases in runoff in the La Plata basin of southern South America, southeastern through central North America, the southeastern quadrant of Africa, and northern Australia. In northern Eurasia and far northwestern North America, the ensemble shows a strong upward trend of runoff, which is consistent with, though more robust than, a general upward trend in the observations in this zone. (Patterns of change in model precipitation (not shown) are qualitatively similar to those of model runoff, as shown previously for similar experiments⁸.)

Differences between ensemble trends and observed trends are also apparent. Outside the tropics, the observed trends show less coherence in space than the ensemble trends. Consistent disagreement in sign of the trend is most apparent in Central America and northern South America (where a marked bias in average runoff has already been noted), northeastern Europe, and central and southeast Asia.

The correlation between ensemble and observed trends in streamflow across basins is +0.34. The correlations between trends computed from individual models and the observed trends are all positive, ranging from +0.05 to +0.28, with a mean value of +0.16. A linear-regression slope of 1.51 for observed versus ensemble

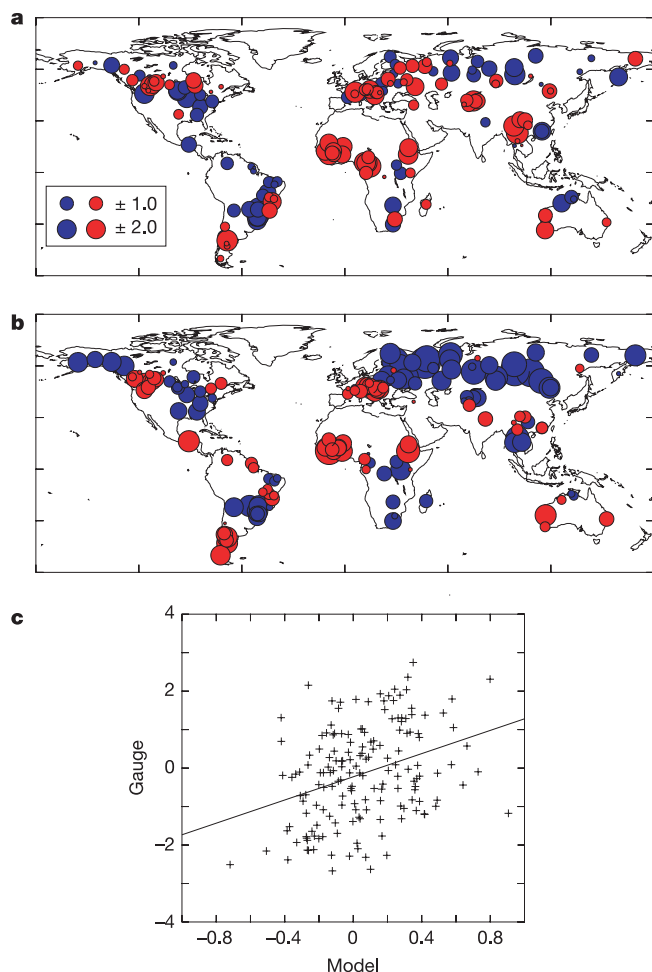


Figure 2 | Global distributions of trend (Z) in streamflow from 1900–70 to 1971–98. a, Stream-gauge observations. b, Ensemble (arithmetic) means of 35 model-run Z values, multiplied by $35^{1/2}$ to account for the reduction in variance caused by averaging. c, Plot of observations against means of 35 model-run Z values. The ordinary least-squares regression line shown has the equation gauge data = $1.51 \times (\text{model ensemble}) - 0.23$.

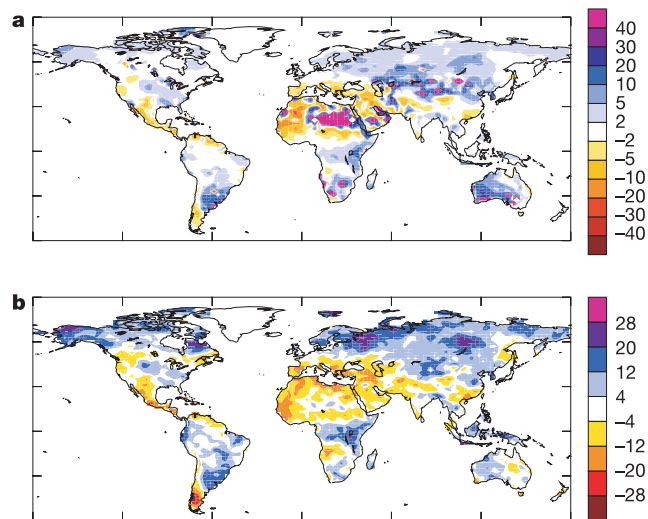


Figure 3 | Relative change in runoff during the twentieth century.

a, Ensemble (arithmetic) mean of relative change (percentage) in runoff for the period 1971–98, computed as 100 times the difference between 1971–98 and 1900–70 runoff in the 20C3M experiments, divided by 1900–70 runoff. **b**, Number of runs (out of a total of 35) showing a positive change minus the number showing a negative change.

trends indicates that observed trends are larger on average than modelled trends, but not significantly so.

Could the positive correlation between the ensemble and observed trends have arisen, by chance, as a result of internal (that is, unforced, natural) variability in the climate system? To address this question of statistical significance, we needed to estimate the sampling distribution of the correlation coefficient between the ensemble trends and the trends computed from repeated realizations of an unforced climate system. To do this, we formed as many distinct 99-year segments of output from the PICNTRL experiment as were available from the 12 models in the ensemble. This sampling yielded 49 synthetic sets of observations, which were mapped to the period 1900–98, masked to allow the use only of values from years and gauges when and where real observations were made, and then used to compute the trend statistics. We determined the correlation of trends in each of these synthetic observation sets with the ensemble average of those from the 20C3M time series. The 49 correlation values ranged from -0.32 to $+0.33$, with a mean of 0.01 ; the 49 regression slopes ranged from -1.12 to $+1.25$, with a mean of 0.05 . We assume that the 49 correlation values can be used to approximate the distribution from which the value $+0.34$ would have been drawn under the null hypothesis of a stationary hydroclimate. Because none of the 49 values are as large as $+0.34$, we infer that the correlation between the forced-model ensemble trends and the observed trends is statistically significant. This inference relies on the assumption that the models faithfully represent interbasin correlation of internal variations of runoff in the models.

Figures 3 and 4 show twentieth-century and twenty-first-century percentage changes in runoff estimated by the model ensemble, with indications of the degree of agreement among models on the direction of change. The model projections for the twenty-first century are dependent on various assumptions, for example those connected with future greenhouse-gas emissions, volcanic activity and solar variability. Quantitative projections by the model ensemble also are affected by large model errors in some basins (Fig. 1), but the demonstrated retrospective skill suggests qualitative validity of the projections. The ensemble-average change in runoff by the period 2041–60 shows a pattern generally consistent with that of twentieth-century change, although amplified and with important qualitative

differences. In general, areas of increased runoff shrink over time (that is, from the late twentieth century to the mid twenty-first century), whereas areas of decreased runoff grow. Initial increases of runoff in the twentieth century are projected to reverse in the twenty-first century in eastern equatorial South America, southern Africa and the western central plains of North America. Modelled drying of the Mediterranean region extends farther north into Europe in the twenty-first-century runs than in the twentieth-century runs.

Almost all model runs agree on the direction of twenty-first-century change in certain regions (Fig. 4). These agreements include increases (typically 10–40% by 2050) in the high latitudes of North America and Eurasia, in the La Plata basin of South America, in eastern equatorial Africa and in some major islands of the equatorial eastern Pacific Ocean. Prominent regions of agreement on decreasing (typically 10–30%) runoff include southern Europe, the Middle East, mid-latitude western North America, and southern Africa.

On the basis of this analysis, it seems that a significant part of twentieth-century hydroclimatic change was externally forced, that larger changes can be expected in the coming decades, and that climate models can help now to characterize future changes. Henceforth it may be prudent to include projections of forced hydroclimatic change as factors in assessments of water availability, thereby facilitating their consideration not only in water management but also in economic and ecological assessment and planning.

METHODS

From a previously defined set of 663 gauged river basins⁹, we selected the 165 gauges judged most suitable for analysis of hydroclimatic change. To be included, a basin was required to have at least 28 years of data and no more than 10% of values missing during the period of record. To avoid overweighting of relatively gauge-rich Europe and North America in the analyses, only basins with drainage area greater than 50,000 km² were included for those continents. Net diversions for irrigation (diversions minus return flows, estimated as the product of irrigated area in the basin¹⁰ and the excess, if any, of mean potential evaporation¹¹ over mean precipitation¹²) were required to be less than 10% of mean flow. Results reported throughout this paper were only slightly sensitive to these subjective numerical constraints on record length, missing values, basin area, and irrigation. The monthly time series of observed discharge were obtained from the Global Runoff Data Centre and averaged to annual values for all analyses reported here. Delineation of the drainage basin associated with each gauge was determined by use of the Simulated Topological Network (STN-30p)¹³.

For year i , the conversion from basin-average model annual runoff $y(i)$ to model streamflow $q(i)$ was given by $q(i) = rq(i-1) + (1-r)y(i)$. For each model and each gauge, the value of r was assigned the difference between the lag-1 autocorrelation of annual values from observed streamflow and the lag-1 autocorrelation of annual model runoff; in the rare cases in which this difference exceeded the largest of all observed values of autocorrelation in the observed discharge (0.90), it was set to the latter value instead. The initial value of q was set equal to the time-average value of y . For comparability, only years with observations were sampled from the models when making comparisons with observations.

The r.m.s. difference between the observed and modelled natural logarithm of discharge ranged from 0.98 to 3.5. Ensembles were formed from the 12 models for which the r.m.s. difference was less than 1.3. In terms of institutional designations used by the Program for Climate Model Diagnosis and Intercomparison (PCMDI), these 12 models are CCSM3, CGCM3.1(T63), ECHAM5/MPI-OM, ECHO-G, FGOALS-g1.0, GFDL-CM2.0, GFDL-CM2.1, GISS-AOM, MIROC3.2(hires), MRI-CGCM2.3.2, UKMO-HadCM3 and UKMO-HadGEM1.

The 1900–98 time range was selected for our analysis because it was the longest for which all models provided output. The choice of 1970 for the break in analyses of change was based partly on our observation from previous model investigations that global-mean measures of hydroclimatic change became noticeable at about this time; additionally, this choice maximized the number of basins with data both before and after the break. We investigated the sensitivity of our results to changes in the break year and found that the results for a 1980 break were similar to those for a 1970 break. A 1960 break generally resulted in smaller (and less significant) trends, and a 1990 break resulted in more variable (and less significant) trends because of the small sampling period thereafter. For break years of 1960, 1970, 1980 and 1990, we found that 12, 0, 0

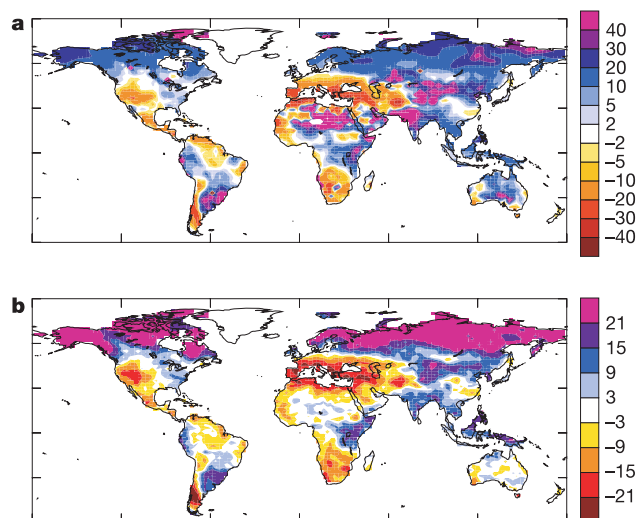


Figure 4 | Relative change in runoff in the twenty-first century.

a, Ensemble (arithmetic) mean of relative change (percentage) in runoff for the period 2041–60, computed as 100 times the difference between 2041–60 runoff in the SRESA1B experiments and 1900–70 runoff in the 20C3M experiments, divided by 1900–70 runoff. **b**, Number of pairs of runs (out of an available total of 24 pairs) showing a positive change minus the number showing a negative change.

and 5, respectively, of the 49 synthetic observations yielded correlations exceeding those of the real observations. If all 21 models (62 runs of 20C3M and 81 segments of PICNTRL) are used instead of the selected 12 models, the numbers of synthetic observations with correlations exceeding those of the real observations are 20, 6, 1 and 7 (out of 81) when using a break year of 1960, 1970, 1980 and 1990.

Received 18 May; accepted 12 October 2005.

1. Reiter, L., Falk, H., Groat, C. & Coussens, C. M. (eds) *From Source Water to Drinking Water: Workshop Summary* (National Academies Press, Washington DC, 2004).
2. United Nations Educational Scientific and Cultural Organization. *Water for People—Water for Life, The United Nations World Water Development Report* (Berghahn Books, Oxford, 2003).
3. Shiklomanov, I. A. & Rodda, J. C. (eds) *World Water Resources at the Beginning of the 21st Century* (Cambridge Univ. Press, Cambridge, 2003).
4. Mooney, H., Cropper, A. & Reid, W. Confronting the human dilemma. *Nature* **434**, 561–562 (2005).
5. Cazenave, A. *et al.* Space techniques used to measure change in terrestrial waters. *Eos* **48**, 59 (2004).
6. Allen, M. & Ingram, W. J. Constraints on future changes in climate and the hydrologic cycle. *Nature* **419**, 224–232 (2002).
7. Brockwell, P. J. & Davis, R. A. *Time Series: Theory and Methods* Ch. 7 (Springer, New York, 1987).
8. Manabe, S., Wetherald, R. T., Milly, P. C. D., Delworth, T. L. & Stouffer, R. J. Century-scale change in water availability: CO₂-quadrupling experiment. *Clim. Change* **64**, 59–76 (2004).
9. Fekete, B. M., Vörösmarty, C. J. & Grabs, W. High-resolution fields of global runoff combining observed river discharge and simulated water balances. *Glob. Biogeochem. Cycles* **16** (2002); published online 7 August 2002 (doi:10.1029/1999GB001254).
10. Siebert, S., Döll, P., Feick, S. & Hoogeveen, J. *Global Map of Irrigated Areas version 2.2* (Johann Wolfgang Goethe University, Frankfurt am Main, Germany; Food and Agriculture Organization of the United Nations, Rome, 2005).
11. Darnell, W. L. *et al.* *Surface Radiation Budget: A Long-term Global Dataset of Shortwave and Longwave Fluxes* [online] (http://www.agu.org/eos_elec/95206e.html) (1996).
12. Adler, R. F. *et al.* The Version 2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeorol.* **4**, 1147–1167 (2003).
13. Vörösmarty, C. J., Fekete, B. M., Meybeck, M. & Lammers, R. B. Global system of rivers: Its role in organizing continental land mass and defining land-to-ocean linkages. *Glob. Biogeochem. Cycles* **14**, 599–621 (2000).

Acknowledgements We thank T. Cohn, T. Delworth, I. Held, G. Hodgkins, H. Lins and R. Stouffer for advice. Streamflow data were provided by national hydrometric services through the Global Runoff Data Centre, Koblenz, Germany. Climate-model outputs were provided by modelling centres through PCMDI. We acknowledge the JSC/CLIVAR Working Group on Coupled Modelling and their Coupled Model Intercomparison Project and Climate Simulation Panel for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. This investigation was performed within the IPCC Global Hydroclimates Analysis Project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.C.D.M. (cmilly@usgs.gov).

Palaeoanatomy and biological affinities of a Cambrian deuterostome (Stylophora)

Sébastien Clausen¹ & Andrew B. Smith²

Stylophora are a peculiar extinct group of asymmetrical deuterostomes whose biological affinity has been fiercely debated^{1–15}. Disarticulated skeletal elements of a ceratocystid stylophoran recovered from the earliest Middle Cambrian of Morocco are not only the oldest stylophorans in the fossil record, but their exceptional preservation provides crucial data on the microstructure of its skeleton. Stylophoran plates are constructed of a three-dimensional mesh, termed ‘stereom’, identical to that of living echinoderms in which stereom microstructure provides a reliable guide to the nature of the investing soft tissues^{16–18}. Using modern echinoderm anatomy to interpret stereom microstructure of stylophoran elements, here we show that the large proximal lumen of their appendage was filled with muscle and that ligamentary tissues bound distal elements firmly together. We find no evidence for a mouth in the proximal lumen and no evidence that the covering plates of the appendage were articulated. Thus, although skeletal structure suggests that stylophorans are echinoderms, their appendage was not a feeding arm but a muscular locomotory organ.

One of the most enduring problems in deuterostome phylogeny has been the position of a peculiar group of asymmetric fossils named Stylophora. For a long time these fossils were considered to be primitive pre-radiate echinoderms with a mobile stem^{1–5} or a single arm^{6–8}. In the 1960s, however, an alternative view arose in which Stylophora were seen as ‘calcichordates’ or primitive chordates still retaining an echinoderm skeleton^{9–12}. A third interpretation has also been advanced that stylophorans, far from being primitive, represent

highly derived echinoderms lying in the echinoderm crown group and sister group to the crinoids^{13–15}.

Stylophorans have an asymmetric body and a single appendage (Fig. 1), and it is in the reconstruction of this appendage that the three rival interpretations most clearly diverge. According to the primitive echinoderm model (Fig. 1b), the stylophoran appendage is a motile stalk that is filled with muscle and locomotory in function^{2,3,5}. In the calcichordate model (Fig. 1c), the appendage is a tail with muscle blocks and a notochord ending proximally at the brain^{10,12}. The crinozoan model (Fig. 1d) has the appendage as a composite structure in which the distal part is an ambulacrum with tube-feet and moveable cover-plates and the proximal region is an oral tegmen housing the mouth and pharynx^{14,15}.

Critical evaluation of these three interpretations has proved difficult owing to the lack of hard data from which to reconstruct soft-tissue anatomy. In modern echinoderms, however, there is a well-established link between skeletal microstructure and the nature of the overlying soft tissue^{16–18}. Stylophorans have a skeleton of

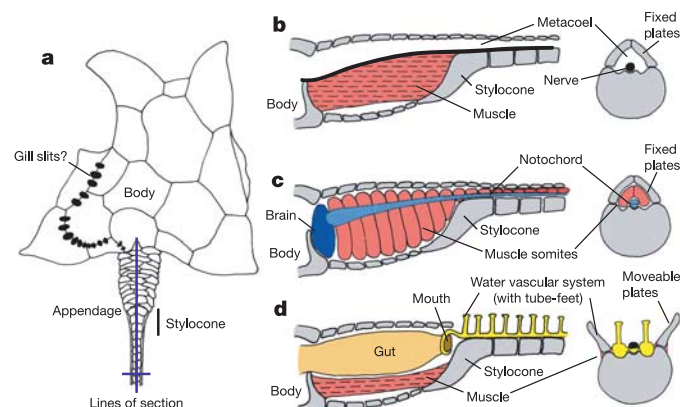


Figure 1 | The stylophoran *Ceratocystis*. **a**, Basic anatomical features. **b–d**, Three current interpretations of the soft-tissue anatomy of the stylophoran appendage in proximal longitudinal (left) and distal transverse (right) section: primitive echinoderm model (**b**), calcichordate model (**c**) and crinozoan model (**d**).

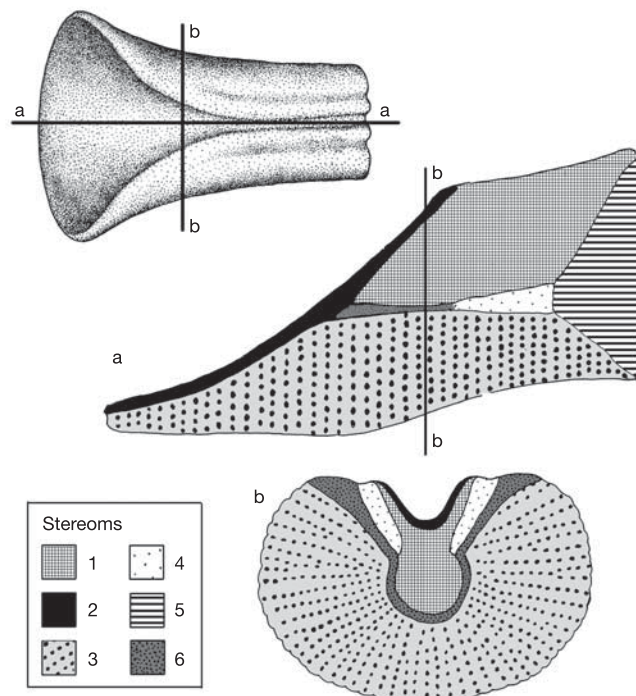


Figure 2 | The stylocone of *Ceratocystis* sp. **a–a** and **b–b** indicate planes of longitudinal (**a**) and transverse (**b**) sections. Numbered stereom types are described in the Supplementary Information.

¹Laboratoire de Paléontologie et Paléogéographie du Paléozoïque, Université des Sciences et Technologies de Lille, F-59655 Villeneuve d'Ascq cedex, France. ²Department of Palaeontology, Natural History Museum, London SW7 5BD, UK.

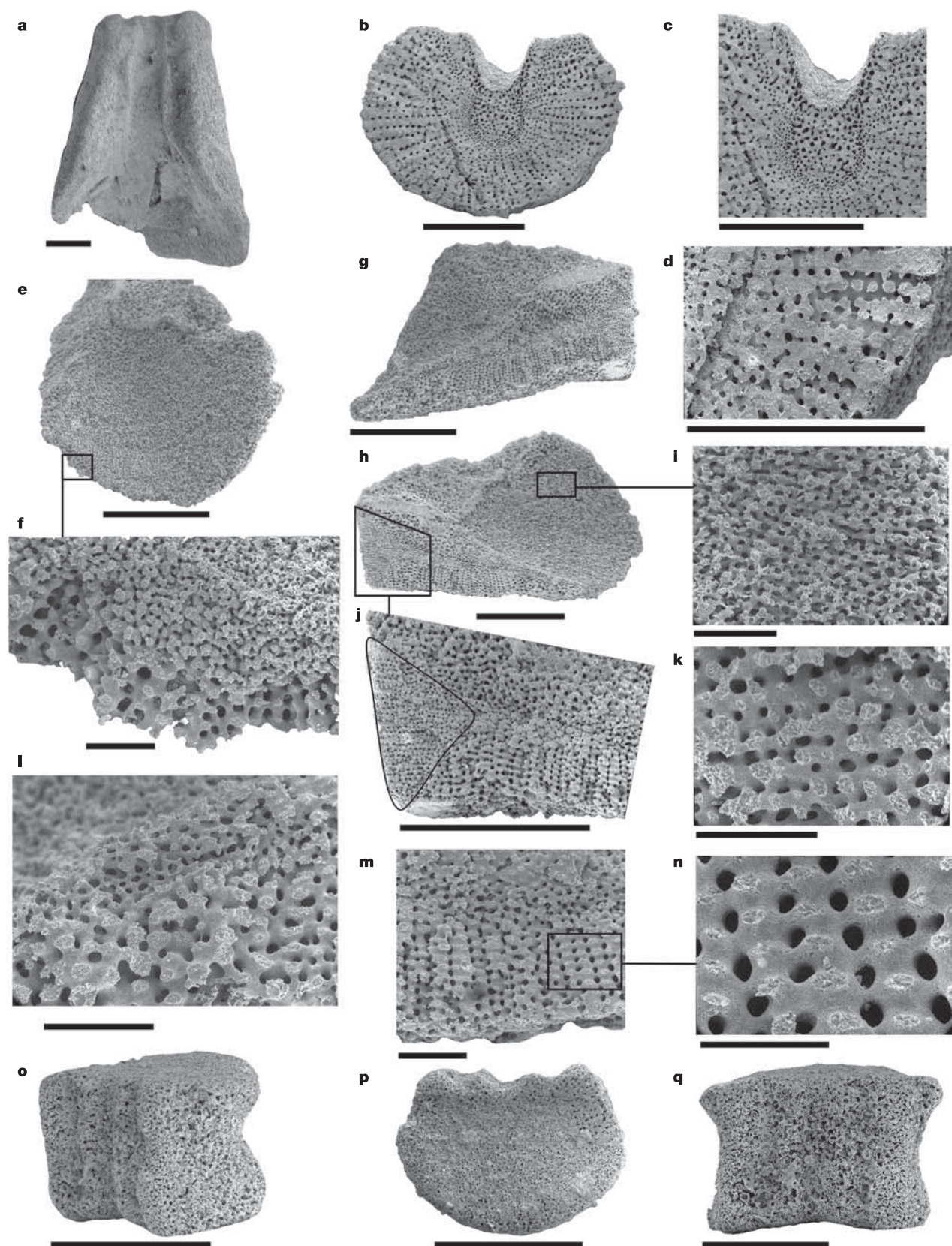


Figure 3 | Stereo of the stylocone and distal appendage ossicle of *Ceratocystis*. Shown are scanning electron microscopy images of stereo of the stylocone (a–n) and distal appendage ossicle (o–q) of *Ceratocystis*. **a**, Upper view, proximal surface to base. **b–d**, Median cross-section (**b**, **c**, part; **d**, counterpart). **e–f**, Proximal face showing type-2 stereo overlying type-3 stereo. **g–n**, Sagittal section (**g**, left half; **h**, right half). **i**, Detail of type-2 labyrinthine stereo from inner face in external view. **j**, Transverse

section towards distal face showing wedge of type-5 galleried stereo. **k**, Detail of type-5 galleried stereo. **l**, Cross-section showing type-2 stereo overlying type-3 stereo. **m**, **n**, Type-3 stereo in cross-section, lower face to bottom. **o–q**, Distal appendage element in oblique lateral view, end view and upper view. Scale bars, 500 μ m (**a–e**, **g**, **h**, **j**, **o–q**); 100 μ m (**f**, **i**, **l**, **m**); 50 μ m (**k**, **n**).

stereom like that of extant echinoderms⁶; thus, in principle it should be possible to use stereom microstructure to reconstruct stylophoran anatomy. The recovery of well-preserved appendage elements belonging to the stylophoran *Ceratocystis* from the earliest Middle Cambrian of Morocco allows, for the first time, such an approach to be taken.

The stylophoran ossicles described here come from the *Micmacca* Breccia of the Lemdad Syncline, western High Atlas, Morocco. The *Micmacca* Breccia consists of volcano and bioclastic bedded and nodular limestones that alternate with shale or sandstone intervals^{19,20}, and encompasses the *Cephalopyge notabilis* (pars) and *Ornamentaspis frequens* biozones (both Tissafinian in age, basal Middle Cambrian according to the Moroccan chronostratigraphic chart²¹). Stylophoran elements recovered include five stylocones, and 12 distal appendage elements, together with a few marginal plates. The ossicles are preserved in three dimensions and the original calcite stereom is replaced by iron oxides. A similar preservation has been observed in the Middle Cambrian of Bornholm, Denmark²².

The *Micmacca* ossicles represent the oldest remains of stylophorans in the fossil record described so far. The most characteristic ossicle of stylophorans is the stylocone, a conical element of highly distinctive morphology situated in the median region of the appendage (Figs 1 and 2). The stylocones from Morocco are unusual in lacking both transverse channels on their upper surface and dorsal spikes or flanges. This combination of features is seen only in the basal stylophoran *Ceratocystis*. Associated thecal plates (plates M₂

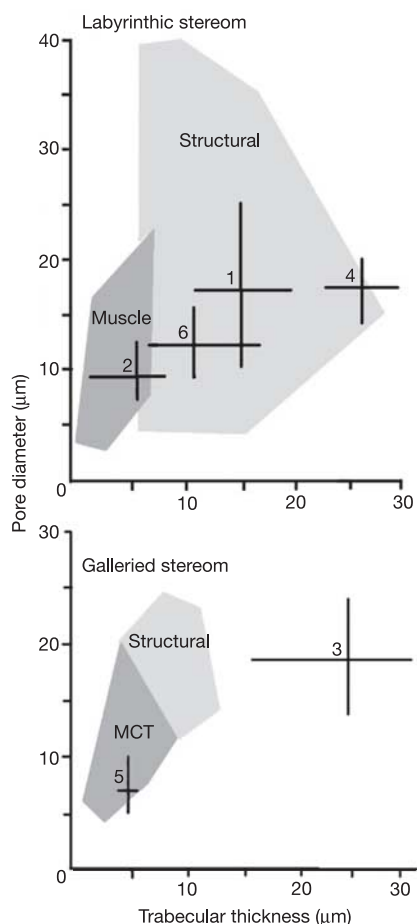


Figure 4 | Stereom fields in modern echinoderms. Dark shaded areas show the stereom fields indicative of muscle and mutable collagenous tissue (MCT) attachment areas for labyrinthine and galleried stereoms taken from ref. 18 (structural stereom fields are shown in a lighter shading). The six stereoms of *Ceratocystis* indicated in Fig. 2 and listed in the Supplementary Information are indicated (range and mean).

and S₂) are also closely comparable to those of *Ceratocystis* (see ref. 23). In the absence of more complete material, however, we assign this fossil only tentatively to *Ceratocystis* (as ?*Ceratocystis* sp.).

Stereom morphologies observed in the plates of ?*Ceratocystis* were compared with an extensive compilation of stereom data taken from modern echinoderms¹⁸. Two forms of stereom are present in ?*Ceratocystis*: labyrinthine stereom, in which pores are irregular in size and show no alignment; and galleried stereom, in which pores are aligned in one direction, usually perpendicular to the outer face (nomenclature follows refs 17, 18). Each can vary in coarseness, and a description of the six stereom types that we distinguish is given in the Supplementary Information. For each stereom morphology, we measured *D* (the average maximum diameter of the pore) and *T* (the average minimum thickness of the calcite bars separating pores).

The stylocone of ?*Ceratocystis* is constructed of six distinct zones of stereom (Figs 2 and 3). Distal and proximal facets of the stylocone of ?*Ceratocystis* have stereoms indicative of different types of articulation (Fig. 4). The flat, distal facet is made up of a galleried stereom (type 5) that in extant echinoderms is always associated with collagen fibres that firmly bind plates together, as in the crinoid stem. The connection between the stylocone and the first distal ossicle was therefore a rigid ligamentary articulation. By contrast, the thin layer of labyrinthine meshwork (type 2) that coats the whole proximal concavity of the stylocone is characteristic of muscle attachment surfaces. The large proximal cavity of the aulacophore therefore housed a massive muscle. Muscle stereom extends right to the border of the median furrow, which folds downward and ends at the proximal notch. The structural labyrinthine stereom (type 1) of the median groove gives no clues about what soft tissue overlay.

The stereom arrangement in more distal elements of the appendage is similar to that of the stylocone except that both proximal and distal faces are composed of medium galleried stereom, indicative of collagen fibres and firm ligamentous binding. We can find no stereom evidence that cover-plates of the appendage were routinely opened and closed. The overlying cover-plates rest along the lateral ridges of these distal ossicles; here, stereom is very different from that typical of articulation surfaces, where stereom at the point of articulation is dense and smooth with low porosity and a band of fine stereom for muscle attachment flanks this region (for example, see edrioasteroids in Fig. 7.1 of ref. 24). In ?*Ceratocystis*, by contrast, a rather open stereom underlies the whole abutment surface. This stereom is coarser than that predicted for muscle and may have been associated with binding ligament, although it lacks a clear galleried structure. A dense labyrinthine stereom is present but confined to the inner ridges and the floor of the lateral grooves, where, according to the crinozoan model, tube-feet should be found. There are no flanking muscle fields.

On the basis of stereom evidence, it is possible to evaluate critically the three current interpretations of the stylophoran appendage. We reject the calcichordate model (Fig. 1c) because the presence of a fine superficial layer of stereom indicates that muscle fibres inserted into stereom rather than being arranged as myotome blocks. The large thecal flanges at the base of the appendage that face the stylocone are almost certainly the anchorage point for this muscle and not the flooring to a brain. The presence of binding ligament between the stylocone and distal elements implies a rigid appendage and argues against the need for notochord and muscle blocks in this part of the appendage.

The crinozoan model (Fig. 1d) is also unlikely. Stereom suggests that cover-plates are immobile and the lateral grooves do not have basins for tube-feet. Furthermore, muscle extends right up to the median furrow on the proximal face of the stylocone, leaving little space for the postulated mouth and pharynx.

Fixed cover-plates and a massive proximal muscle would, however, fit with the third interpretation: namely, the appendage as a muscular stem (Fig. 1b). Although no extant stalked echinoderm has extensive muscle in its stem, the stalk of pterobranchs, sister taxon to the

echinoderms, is muscular and locomotory in function²⁵, and it is possible that the proximal stem of the extinct pleurocystitid cystoids also contained muscle²⁶. Stereom microstructure provides convincing evidence that stylophorans had muscle and ligament attachment similar to that of modern echinoderms and a muscular locomotory appendage. Together with their complete lack of radial symmetry, this evidence supports their interpretation as stem-group echinoderms.

Received 15 June; accepted 3 August 2005.

1. Bather, F. A. A class of Echinoderma without trace of radiate symmetry. *Arch. Zool. Ital.* **14**, 431–439 (1930).
2. Philip, G. M. Carpoids—echinoderms or chordates? *Biol. Rev.* **54**, 439–471 (1979).
3. Kolata, D. R., Frest, T. J. & Mapes, R. H. The youngest carpod: occurrence, affinities and life mode of a Pennsylvanian (Morrowan) mitrate from Oklahoma. *J. Paleontol.* **65**, 844–855 (1991).
4. Gee, H. *Before the Backbone: Views on the Origin of the Vertebrates* (Chapman & Hall, London, 1996).
5. Smith, A. B. The preradial history of echinoderms. *Geol. J.* **40**, 255–280 (2004).
6. Ubaghs, G. in *Treatise on Invertebrate Paleontology part 5 Echinodermata 1* (ed. Moore, R. C.) 495–565 (Geological Society of America and Univ. Kansas Press, Boulder, Colorado, 1968).
7. Parsley, R. L. in *Echinoderm Phylogeny and Evolutionary Biology* (eds Paul, C. R. C. & Smith, A. B.) 345–361 (Clarendon, Oxford, 1988).
8. Sprinkle, J. in *Origin and Early Evolution of Metazoa* (eds Lipps, J. H. & Signor, P. W.) 375–398 (Plenum, New York, 1992).
9. Jefferies, R. P. S. Some fossil chordates with echinoderm affinities. *Symp. Zool. Soc. Lond.* **20**, 163–208 (1967).
10. Jefferies, R. P. S. *The Ancestry of the Vertebrates* (British Museum, Natural History, London, 1986).
11. Cripps, A. P. A cladistic analysis of the cornutes (stem chordates). *Zool. J. Linn. Soc.* **102**, 333–366 (1991).
12. Jefferies, R. P. S., Brown, N. A. & Daley, P. E. J. The early phylogeny of chordates and echinoderms and the origin of chordate left-right asymmetry and bilateral symmetry. *Acta Zool.* **77**, 101–122 (1996).
13. Sumrall, C. D. The role of fossils in the phylogenetic reconstruction of Echinodermata. *Paleontol. Soc. Pap.* **3**, 267–288 (1997).
14. David, B., Lefebvre, B., Mooi, R. & Parsley, R. Are homalozoans echinoderms? An answer from the extraxial–axial theory. *Paleobiology* **26**, 529–555 (2000).
15. Lefebvre, B. Functional morphology of stylophoran echinoderms. *Palaeontology* **46**, 511–555 (2003).
16. Roux, M. Microstructural analysis of the crinoid stem. *Paleontol. Contrib. Univ. Kansas* **75**, 1–7 (1975).
17. Smith, A. B. Stereom microstructure of the echinoid test. *Spec. Pap. Paleontol.* **25**, 1–81 (1980).
18. Smith, A. B. in *Skeletal Biomineralization: Patterns, Processes, and Evolutionary Trends* (ed. Carter, J. G.) 413–443 (Van Nostrand Reinhold, New York, 1990).
19. Álvaro, J. J. *The Lower-Middle Cambrian Transition in the Western Mediterranean Region: Biodiversity and Paleogeographic Patterns*. Thesis, Univ. Lille 1 (2002).
20. Clausen, S. *The Non-reefal Echinoderm-sponge Meadows of the Lower-Middle Cambrian Transition in the Western Mediterranean Region*. Thesis, Univ. Lille 1 (2004).
21. Geyer, G. & Landing, E. The Cambrian of the Moroccan Atlas regions. *Beringeria* **2** (special issue), 7–46 (1995).
22. Berg-Madsen, V. Middle Cambrian cystoid (*sensu lato*) columnals from Bornholm. *Denmark. Lethaia* **19**, 67–80 (1986).
23. Ubaghs, G. Le genre *Ceratocystis* Jaekel (Echinodermata, Stylophora). *Paleontol. Contrib. Univ. Kansas* **30**, 1–16 (1967).
24. Sumrall, C. D. & Bowsher, A. L. *Gigantoclavus*, a new genus of Pennsylvanian edrioasteroid from North America. *J. Paleontol.* **70**, 986–993 (1996).
25. Benito, J. & Pardos, F. in *Microscopic Anatomy of Invertebrates* vol. 15 *Hemichordata, Chaetognatha, and the Invertebrate Chordates* (eds Harrison, F. W. & Ruppert, E. E.) 15–102 (Wiley-Liss, New York, 1997).
26. Brower, J. C. A new pleurocystitid rhombiferan echinoderm from the Middle Ordovician Galena Group of northern Iowa and southern Minnesota. *J. Paleontol.* **73**, 129–153 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank J. J. Álvaro for comments on an earlier draft of this paper. Funding was provided by High Lat resources through the European Commission's Access to Research Infrastructure action of the Improving Human Potential Programme and the Swedish Museum of Natural History.

Author Contributions The two authors contributed equally to this paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.C. (sebastien.clausen@ed.univ-lille1.fr).

Superspreading and the effect of individual variation on disease emergence

J. O. Lloyd-Smith^{1,2}, S. J. Schreiber³, P. E. Kopp⁴ & W. M. Getz¹

Population-level analyses often use average quantities to describe heterogeneous systems, particularly when variation does not arise from identifiable groups^{1,2}. A prominent example, central to our current understanding of epidemic spread, is the basic reproductive number, R_0 , which is defined as the mean number of infections caused by an infected individual in a susceptible population^{3,4}. Population estimates of R_0 can obscure considerable individual variation in infectiousness, as highlighted during the global emergence of severe acute respiratory syndrome (SARS) by numerous 'superspreading events' in which certain individuals infected unusually large numbers of secondary cases^{5–10}. For diseases transmitted by non-sexual direct contacts, such as SARS or smallpox, individual variation is difficult to measure empirically, and thus its importance for outbreak dynamics has been unclear^{2,10,11}. Here we present an integrated theoretical and statistical analysis of the influence of individual variation in infectiousness on disease emergence. Using contact tracing data from eight directly transmitted diseases, we show that the distribution of individual infectiousness around R_0 is often highly skewed. Model predictions accounting for this variation differ sharply from average-based approaches, with disease extinction more likely and outbreaks rarer but more explosive. Using these models, we explore implications for outbreak control, showing that individual-specific control measures outperform population-wide measures. Moreover, the dramatic improvements achieved through targeted control policies emphasize the need to identify predictive correlates of higher infectiousness. Our findings indicate that superspreading is a normal feature of disease spread, and to frame ongoing discussion we propose a rigorous definition for superspreading events and a method to predict their frequency.

For sexually transmitted and vector-borne diseases, host contact rates have long served as surrogates for individual infectiousness^{3,12–14}, leading to the assertion of a general '20/80 rule' (whereby 20% of cases cause 80% of transmission¹³) and to the influential concept of high-risk 'core groups'^{3,12,13}. For directly transmitted infections, however, the overall infectiousness of each case—that is, the number of other individuals infected during the infectious lifetime of a single individual—arises from a complex mixture of host, pathogen and environmental factors (see Supplementary Notes). Consequently, the degree of infectiousness is distributed continuously in any population^{4,7,11,15,16} and, crucially, distinct risk groups often cannot be defined *a priori*^{2,11}. This impedes the conventional approach to adding heterogeneity to epidemic models, in which populations are divided into homogeneous subgroups^{2–4,17}. Research on continuous individual variation in infectiousness for directly transmitted infections has been largely restricted to within-household transmission^{18,19}, or to variation in infectious period^{20,21} or social network²². Some recent studies have used contact tracing

data to investigate specific questions in light of observed variation^{8,16}, but a broad understanding of the role of individual variation in outbreak dynamics is lacking.

As a theoretical basis, we introduce the 'individual reproductive number', ν , as a random variable representing the expected number of secondary cases caused by a particular infected individual. Values for ν are drawn from a continuous probability distribution with population mean R_0 that encodes all variation in infectious histories of individuals, including properties of the host and pathogen and environmental circumstances. In this framework, superspreading events (SSEs) are not exceptional events⁹, but important realizations from the right-hand tail of a distribution of ν (refs 7, 15). Stochastic effects in transmission are modelled using a Poisson process⁴, so that the number of secondary infections caused by each case, Z , is described by an 'offspring distribution' $\Pr(Z = k)$ where $Z \sim \text{Poisson}(\nu)$.

By considering three possible distributions of ν , we generate three candidate models for the offspring distribution: (1) in generation-based models neglecting individual variation, $\nu = R_0$ for all cases, yielding $Z \sim \text{Poisson}(R_0)$; (2) in differential-equation models with homogeneous transmission and constant recovery rates, ν is exponentially distributed, yielding $Z \sim \text{geometric}(R_0)$; (3) in a more general formulation, we let ν be gamma-distributed with mean R_0 and dispersion parameter k , yielding $Z \sim \text{negative binomial}(R_0, k)$ (ref. 23). The negative binomial model includes the conventional Poisson ($k \rightarrow \infty$) and geometric ($k = 1$) models as special cases. It has variance $R_0(1 + R_0/k)$, so smaller values of k indicate greater heterogeneity.

We gathered empirical offspring distributions from detailed contact tracing or surveillance data sets, and challenged the candidate models using model selection techniques²⁴ (see Supplementary Notes). For SARS outbreaks in Singapore and Beijing, the negative binomial model is unequivocally favoured (Fig. 1a and Supplementary Table 1). Conventional models assuming homogeneity cannot reproduce the observed transmission patterns. For the Singapore outbreak, the maximum-likelihood estimate $\hat{k}$ is 0.16 (90% confidence interval 0.11–0.64), indicating an underlying distribution of ν that is highly overdispersed (Fig. 1a, inset). According to this analysis, the great majority of SARS cases in Singapore were barely infectious (73% had $\nu < 1$) but a small proportion were highly infectious (6% had $\nu > 8$). This is consistent with field reports from SARS-afflicted regions^{5,6} but contrasts with published SARS models^{9,10,25,26}.

Comparing results for eight directly transmitted infections reveals the differing degree of individual variation among diseases and outbreak settings (Fig. 1b, c and Supplementary Tables 1, 2). The Poisson offspring distribution is almost always strongly rejected. The geometric model has considerable support for several data sets, which indicates significant individual variation in transmission

¹Department of Environmental Science, Policy and Management, 140 Mulford Hall, University of California, Berkeley, California 94720-3114, USA. ²Biophysics Graduate Group, University of California, Berkeley, California 94720-3200, USA. ³Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23187-8975, USA.

⁴Centre for Mathematics, University of Hull, Hull HU6 7RX, UK.

rates because real infectious periods are less dispersed than the exponential distribution^{20,21}. The negative binomial model is selected decisively for several data sets, and enables comparative study of diseases through the dispersion parameter. Like SARS, measles in highly vaccinated populations shows high variation in two surveillance data sets, with narrow confidence intervals excluding the conventional models (note that heterogeneous vaccination coverage is an important environmental factor contributing to this pattern). Monkeypox and smallpox viruses show intermediate variation, consistent across multiple data sets, and pneumonic plague transmission is slightly less variable. Data limitations prevent definitive conclusions for other diseases. Comparing our findings to the 20/80 rule proposed for sexually transmitted and vector-borne diseases¹³, no general rule emerges but the core principle of heterogeneous transmission is certainly supported (Fig. 1c).

Numerous reports of superspreading events provide further evidence for variation in ν . We reviewed 37 published accounts of SSEs for 11 directly transmitted infections (Fig. 1d; see Supplementary Notes). Unrecognized or misdiagnosed illness is the most common cause of these SSEs, followed by alternative modes of spread (especially airborne), high contact rates, and co-infections that aid transmission. High pathogen load or shedding rates are occasionally implicated, but are rarely measured. A consistent and general definition of SSEs is currently lacking—for SARS, an SSE has been

arbitrarily defined as $Z \geq 8$ (ref. 6), $Z \geq 10$ (ref. 5), $Z > 10$ (ref. 26) or ‘many more than the average number’⁹, and different thresholds are surely needed for measles ($R_0 \sim 11\text{--}18$; ref. 3) or monkeypox ($R_0 < 1$).

We propose this general protocol for defining a superspreading event: (1) estimate the effective reproductive number, R , for the disease and population in question; (2) construct a Poisson distribution with mean R , representing the expected range of Z due to stochasticity without individual variation; (3) define an SSE as any infected individual who infects more than $Z^{(n)}$ others, where $Z^{(n)}$ is the n th percentile of the $\text{Poisson}(R)$ distribution. A 99th-percentile SSE is then any case causing more infections than would occur in 99% of infectious histories in a homogeneous population (Fig. 1d). This approach complements *a priori* identification of potential superspreaders when that is feasible, as for sexually transmitted diseases (where promiscuity drives risk)^{3,12}. In addition, the definition enables prediction of the frequency of SSEs once R_0 and k have been estimated (Supplementary Fig. 1)—an outstanding challenge in emerging disease epidemiology^{8,9}.

To assess the effect of individual variation on disease outbreaks, we analyse a branching process model with negative binomial offspring distribution, corresponding to gamma-distributed ν (Fig. 2a; see Supplementary Notes). Of primary interest is the probability of stochastic extinction, q , after the introduction of a single infected

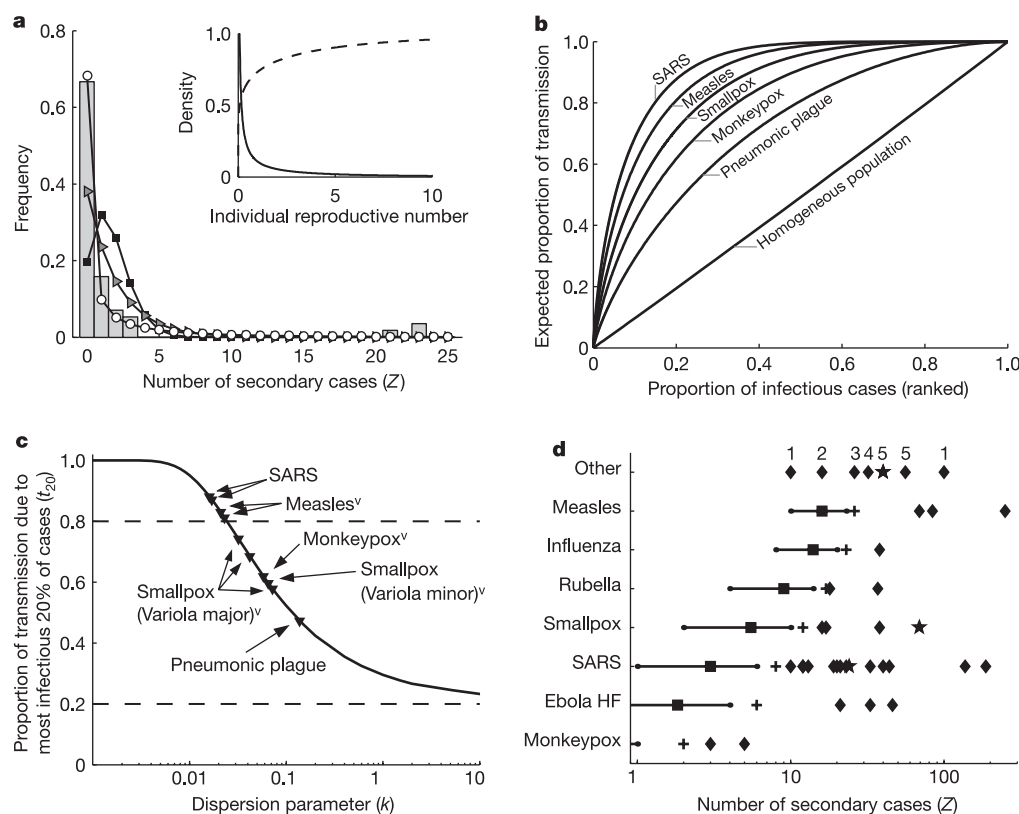


Figure 1 | Evidence for variation in individual reproductive number ν .

a, Transmission data from the SARS outbreak in Singapore in 2003 (ref. 5). Bars show observed frequency of Z , the number of individuals infected by each case. Lines show maximum-likelihood fits for $Z \sim \text{Poisson}$ (squares), $Z \sim \text{geometric}$ (triangles), and $Z \sim \text{negative binomial}$ (circles). Inset, probability density function (solid) and cumulative distribution function (dashed) for gamma-distributed ν (corresponding to $Z \sim \text{negative binomial}$) estimated from Singapore SARS data. **b**, Expected proportion of transmission due to a given proportion of infectious cases, where cases are ranked by infectiousness. For a homogeneous population (all $\nu = R_0$), this relation is linear. For five directly transmitted infections (based on k values in Supplementary Table 1), the line is concave owing to variation

in ν . **c**, Proportion of transmission expected from the most infectious 20% of cases, for 10 outbreak or surveillance data sets (triangles). Dashed lines show proportions expected under the 20/80 rule (top) and homogeneity (bottom). Superscript ‘v’ indicates a partially vaccinated population.

d, Reported superspreading events (SSEs; diamonds) relative to estimated reproductive number R (squares) for twelve directly transmitted infections. Lines show 5–95 percentile range of $Z \sim \text{Poisson}(R)$, and crosses show the 99th-percentile proposed as threshold for SSEs. Stars represent SSEs caused by more than one source case. ‘Other’ diseases are: 1, Streptococcus group A; 2, Lassa fever; 3, Mycoplasma pneumonia; 4, pneumonic plague; 5, tuberculosis. R is not shown for ‘other’ diseases, and is off-scale for monkeypox. See Supplementary Notes for details.

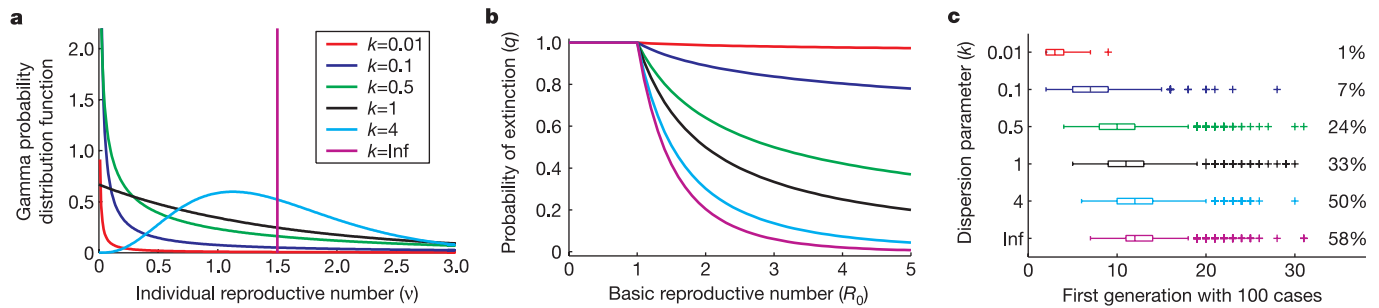


Figure 2 | Outbreak dynamics with different degrees of individual variation in infectiousness. **a**, The individual reproductive number ν is drawn from a gamma distribution with mean R_0 and dispersion parameter k . Probability density functions are shown for six gamma distributions with $R_0 = 1.5$ ($k = \text{Inf}$ indicates $k \rightarrow \infty$). **b**, Probability of stochastic extinction of an outbreak, q , versus population-average reproductive number, R_0 , following introduction of a single infected individual. The value of k increases from

individual (Fig. 2b). For $R_0 < 1$, all invasions die out, as in standard models. For $R_0 > 1$, increased variation strongly favours extinction⁸. For example, if $R_0 = 3$ then $q = 0.06$ under the assumption of homogeneous ν ($k \rightarrow \infty$), or $q = 0.33$ if $k = 1$, but if $k = 0.16$ (as estimated for SARS) then $q = 0.76$. Extinction risk rises owing to a higher proportion of non-transmitting cases when ν is overdispersed (Figs 1a, 2a and Supplementary Fig. 2a). This effect thwarts invasion by diseases that are very potent spreaders on average: for arbitrarily high R_0 , $q \rightarrow 1$ as $k \rightarrow 0$ (Supplementary Fig. 2b). The expected number of cases before extinction is hardly affected by k (Supplementary Fig. 2c), because low- k outbreaks that fail probably lacked SSEs and thus resemble homogeneous outbreaks with lower R_0 . Accordingly, when individual variation is large, extinction occurs rapidly or not at all (Supplementary Fig. 2d).

For outbreaks avoiding stochastic extinction, epidemic growth

rates strongly depend on variation in ν (Fig. 2c and Supplementary Fig. 2e, f). Diseases with high individual variation show infrequent but explosive epidemics after introduction of a single case. This pattern recalls SARS in 2003, for which many settings experienced no epidemic despite unprotected exposure to SARS cases^{27,28}, whereas a few cities suffered explosive outbreaks^{8–10,15,26}. Our results, using $\hat{k} = 0.16$ for SARS, explain this simply by the presence or absence of high- ν individuals in the early generations of each outbreak⁶. In contrast, conventional models (with $k = 1$ or $k \rightarrow \infty$) cannot simultaneously generate frequent failed invasions and rapid growth rates without additional, subjective model structure.

Disease control interventions could increase or decrease individual variation in infectiousness. Infected individuals might reduce their number of non-essential contacts, or governments might impose quarantine or isolation on particular individuals. Here we explore

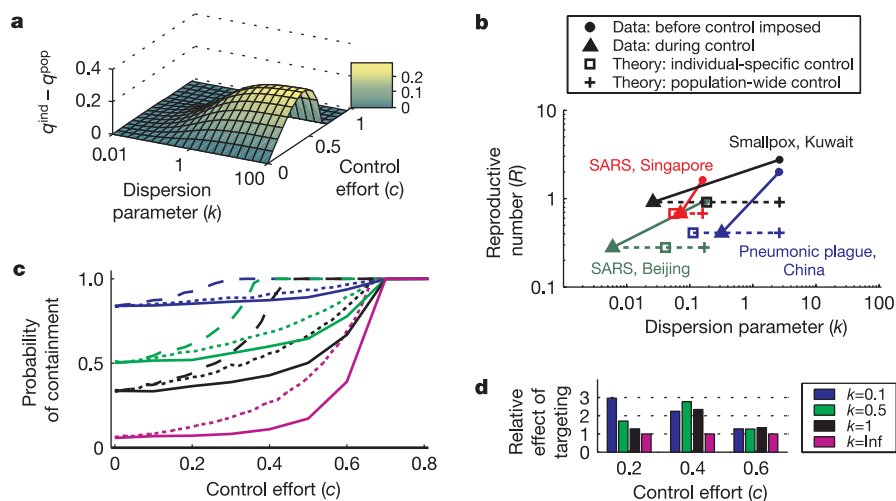


Figure 3 | Implications for control measures. **a**, Increase in extinction probability ($q_{\text{ind}} - q_{\text{pop}}$) under individual-specific control compared to population-wide control, for diseases with $R_0 = 3$ and different degrees of individual variation, k , subject to control effort c . With population-wide control, the infectiousness of all individuals is reduced by a factor c . With individual-specific control, a proportion c of infectious individuals (selected at random) have their infectiousness reduced to zero. The outbreak is assumed to begin with one case, with control present from the outset. **b**, Estimates of $\hat{R}$ and $\hat{k}$ from outbreak data sets before and after control measures were initiated (joined by solid lines; Supplementary Table 2), and post-control values of k_c estimated from theoretical models of control as described in the Supplementary Notes. **c**, Effect of random versus targeted control measures. The probability of outbreak containment (defined as

never reaching the 100-case threshold) for four diseases with $R_0 = 3$ and $k = 0.1$ (blue), $k = 0.5$ (green), $k = 1$ (black) or $k \rightarrow \infty$ (purple). Control policies are population-wide (solid lines), random individual-specific (dotted lines), or targeted individual-specific (dashed lines, where half of all control effort is focused on the most infectious 20% of cases). For $k \rightarrow \infty$, all individuals are identical, so targeting has no effect and dotted and dashed lines overlay one another. **d**, The factor by which targeting increases the effect of control on preventing a major outbreak, relative to random individual-specific control (see Supplementary Notes), when 20%, 40% or 60% of the total population is controlled. Results in **c** and **d** are the mean of 10,000 simulations, with control beginning in the second generation of cases.

several idealized cases theoretically, for an outbreak with offspring distribution $Z \sim$ negative binomial(R_0, k) before control (see Supplementary Notes). Consider the effect of control effort c , where $c = 0$ reflects no control and $c = 1$ reflects complete blockage of transmission. Under population-wide control, the infectiousness of every individual in the population is reduced by a factor c (that is, $\nu_c^{\text{pop}} = (1 - c)\nu$ for all individuals). Under random, individual-specific control, a proportion c of infected individuals (chosen at random) is traced and isolated completely such that they cause zero infections (that is $\nu_c^{\text{ind}} = 0$ for a proportion c of infected individuals, and $\nu_c^{\text{ind}} = \nu$ for the rest). Individual-specific control raises the degree of heterogeneity in the outbreak as measured by the variance-to-mean ratio of Z , whereas population-wide control reduces heterogeneity. Both approaches yield effective reproductive number $R = (1 - c)R_0$, so the threshold control effort for guaranteed disease extinction is $c \geq 1 - 1/R_0$ as in conventional models. For intermediate values of c , however, the individual-specific approach always works better (Fig. 3a and Supplementary Fig. 3a, b), consistent with our finding that higher variation favours disease extinction (Fig. 2b). Branching process theory confirms that $q^{\text{ind}} > q^{\text{pop}}$ whenever $c \in (0, 1 - 1/R_0)$ (see Supplementary Notes).

To assess the realism of these idealized control scenarios, we analysed contact tracing data from four outbreaks before and after imposition of control measures. Control always lowered the estimated dispersion parameter (that is $\hat{k}_c < \hat{k}$) as predicted by the individual-specific model (Fig. 3b), although small sample sizes often led to overlapping confidence intervals (Supplementary Table 2). This increased skew in transmission arose chiefly from undiagnosed or misdiagnosed individuals, who continued to infect others (and even cause SSEs), whereas controlled individuals infected very few. To further examine our control theories, we calculated k_c^{pop} and k_c^{ind} for each data set; $\hat{k}_c$ was always closer to k_c^{ind} , although twice it fell between the two predictions, indicating a possible combination of control mechanisms (Fig. 3b). Real-world control thus seems to increase individual variation, favouring extinction but risking ongoing SSEs. Larger data sets are needed to establish this pattern definitively.

If highly infectious individuals can be identified predictively (see Supplementary Notes) then the efficiency of control could be greatly increased (Fig. 3c, d). Focusing half of all control effort on the most infectious 20% of cases is up to threefold more effective than random control (Fig. 3d). When $k = 0.1$ or 0.5 , outbreak containment is assured for targeted control levels at roughly half the threshold level of $c = 1 - 1/R_0$ for random control. Gains in efficiency increase with more intense targeting of high- ν cases, but saturate as overall coverage c increases (Supplementary Fig. 3c, d). Again, branching process theory generalizes these findings: for a given proportion c of individuals controlled, greater targeting of higher- ν individuals leads to lower effective reproductive number R and higher extinction probability q (see Supplementary Notes).

The data sets analysed here were collected from published literature, and may be subject to selection bias for successful invasions and SSEs rather than typical disease behaviour. Surveillance data sets are less vulnerable to this bias, but may under-report isolated cases. We urge that detailed transmission tracing data be collected and made public whenever possible, even if unexceptional. At a minimum, we propose a new measure for inclusion in outbreak reports: the proportion of cases not transmitting (p_0), which, together with R_0 is sufficient to estimate the degree of variation in ν (Supplementary Fig. 4). As more data become available, trends may emerge in the degree of variation present, for example, for different modes of spread or levels of virulence. Richer data sets may also enable testing of the branching process assumption that case outcomes are independent and identically distributed, by detecting possible correlations in ν values within transmission lineages or systematic changes as outbreaks progress.

Our results have broad implications for emerging disease

epidemiology, and open challenges for further work. Explosive epidemics demand rapid action by authorities and can strain health infrastructures. High extinction probabilities indicate that disease introductions or host species jumps may be more frequent than currently suspected. Cluster-size surveillance for pathogen adaptation²⁹ or dwindling population immunity³⁰ should be tuned to observed levels of variation. Realization of targeted control measures requires a better understanding of factors determining individual infectiousness. This work must be integrated with established theory of sexually transmitted diseases and social networks, where high-risk groups exert nonlinear influence on R_0 because contact rates affect infectiousness and susceptibility equally^{3,4,12,13,22}. All diseases probably show intermediate degrees of covariation between infectiousness and susceptibility, a topic demanding empirical and theoretical study¹⁷. The central role of R_0 in epidemic analysis is unassailable, but our findings show that emerging disease outbreaks cannot be fully understood if individual variation in infectiousness is neglected. Examination of other population processes dependent on small numbers of individuals may yield similar insights.

METHODS

Analysis of disease data. For data sets including the full distribution of Z , we estimated $\hat{R}_0$ and $\hat{k}$ using maximum-likelihood methods. The candidate models were compared using Akaike's information criterion (AIC_c) modified for small sample size. Confidence intervals for $\hat{k}$ were estimated by bias-corrected non-parametric bootstrapping and corroborated by four other methods. For data sets including only estimates of $\hat{R}_0$ and the proportion of cases not transmitting ($\hat{p}_0$), we estimated $\hat{k}$ by solving $\hat{p}_0 = (1 + \hat{R}_0/k)^{-k}$ numerically, and evaluated the candidate models using confidence intervals calculated by two methods. Expected proportions of transmission due to particular groups of infectious individuals (Fig. 1b, c) were calculated using the gamma distribution of ν with estimated values of $\hat{R}_0$ and $\hat{k}$. See Supplementary Notes for details, and for descriptions of data sets.

Branching process analysis. Analysis of branching process models centres on the probability generating function (pgf) of the offspring distribution, $g(s) = \sum_{j=0}^{\infty} \Pr(Z=j)s^j$, defined for $|s| \leq 1$. When $R_0 > 1$, the long-term probability of disease extinction after introduction of a single infected individual is the unique solution of $q = g(q)$ on the interval $(0, 1)$. For a negative binomial offspring distribution $Z \approx \text{NegB}(R_0, k)$, the pgf is $g(s) = (1 + \frac{R_0}{k}(1-s))^{-k}$. Under population-wide control, $Z_c^{\text{pop}} \approx \text{NegB}((1-c)R_0, k)$ and therefore $g_{\text{pop}}(s) = (1 + (1-c)\frac{R_0}{k}(1-s))^{-k}$, and the variance-to-mean ratio is $1 + (1-c)R_0/k$. Under random individual-specific control, the exact pgf is $g_{\text{ind}}(s) = c + (1-c)(1 + \frac{R_0}{k}(1-s))^{-k}$ with variance-to-mean ratio $1 + R_0/k + cR_0$. This scenario can be approximated by $Z_c^{\text{ind,NB}} \approx \text{NegB}((1-c)R_0, k_c^{\text{ind}})$, where k_c^{ind} is the solution to $p_0 + c(1-p_0) = (1 + R_c^{\text{ind}}/k_c^{\text{ind}})^{-k_c^{\text{ind}}}$ and decreases monotonically as c increases. Further details, descriptions of outbreak simulations and formal analysis of control measures are found in the Supplementary Notes.

Received 1 March; accepted 22 August 2005.

- Levin, S. A., Grenfell, B., Hastings, A. & Perelson, A. S. Mathematical and computational challenges in population biology and ecosystems science. *Science* **275**, 334–343 (1997).
- Becker, N. G. & Britton, T. Statistical studies of infectious disease incidence. *J. R. Stat. Soc. B* **61**, 287–307 (1999).
- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, 1991).
- Diekmann, O. & Heesterbeek, J. A. P. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis, and Interpretation* (Wiley, Chichester, 2000).
- Leo, Y. S. et al. Severe acute respiratory syndrome—Singapore, 2003. *Morbidity Mortal. Wkly Rep.* **52**, 405–411 (2003).
- Shen, Z. et al. Superspreading SARS events, Beijing, 2003. *Emerg. Infect. Dis.* **10**, 256–260 (2004).
- Dye, C. & Gay, N. Modeling the SARS epidemic. *Science* **300**, 1884–1885 (2003).
- Lipsitch, M. et al. Transmission dynamics and control of severe acute respiratory syndrome. *Science* **300**, 1966–1970 (2003).
- Riley, S. et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* **300**, 1961–1966 (2003).
- Bauch, C. T., Lloyd-Smith, J. O., Coffee, M. & Galvani, A. P. Dynamically modeling SARS and other newly-emerging respiratory illnesses: past, present, future. *Epidemiology* **16**, 791–801 (2005).

11. Koopman, J. Modeling infection transmission. *Annu. Rev. Public Health* **25**, 303–326 (2004).
12. May, R. M. & Anderson, R. M. Transmission dynamics of HIV infection. *Nature* **326**, 137–142 (1987).
13. Woolhouse, M. E. J. *et al.* Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proc. Natl Acad. Sci. USA* **94**, 338–342 (1997).
14. Lloyd-Smith, J. O., Getz, W. M. & Westerhoff, H. V. Frequency-dependent incidence in models of sexually transmitted diseases: portrayal of pair-based transmission and effects of illness on contact behaviour. *Proc. R. Soc. Lond. B* **271**, 625–634 (2004).
15. Anderson, R. M. *et al.* Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. *Phil. Trans. R. Soc. Lond. B* **359**, 1091–1105 (2004).
16. Gani, R. & Leach, S. Epidemiologic determinants for modeling pneumonic plague outbreaks. *Emerg. Infect. Dis.* **10**, 608–614 (2004).
17. Becker, N. & Marschner, I. in *Stochastic Processes in Epidemic Theory* (eds Picard, P., Gabriel, J. P. & Lefevre, C.) 90–103 (Springer-Verlag, New York, 1990).
18. Becker, N. *Analysis of Infectious Disease Data* 48–59 (Chapman & Hall, London, 1989).
19. Bailey, N. T. J. *The Mathematical Theory of Infectious Diseases* 2nd edn, 263–265 (Griffin, London, 1975).
20. Keeling, M. J. & Grenfell, B. T. Disease extinction and community size: Modeling the persistence of measles. *Science* **275**, 65–67 (1997).
21. Lloyd, A. L. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Proc. R. Soc. Lond. B* **268**, 985–993 (2001).
22. Meyers, L. A., Pourbohloul, B., Newman, M. E. J., Skowronski, D. M. & Brunham, R. C. Network theory and SARS: predicting outbreak diversity. *J. Theor. Biol.* **232**, 71–81 (2005).
23. Boswell, M. T. & Patil, G. P. *Random Counts in Models and Structures* (ed. Patil, G. P.) 7–8 (Pennsylvania State Univ. Press, University Park, 1970).
24. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, New York, 2002).
25. Lloyd-Smith, J. O., Galvani, A. P. & Getz, W. M. Curtailing transmission of severe acute respiratory syndrome within a community and its hospital. *Proc. R. Soc. Lond. B* **270**, 1979–1989 (2003).
26. Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516 (2004).
27. Ha, L. D. *et al.* Lack of SARS transmission among public hospital workers, Vietnam. *Emerg. Infect. Dis.* **10**, 265–268 (2004).
28. Park, B. J. *et al.* Lack of SARS transmission among healthcare workers, United States. *Emerg. Infect. Dis.* **10**, 244–248 (2004).
29. Ferguson, N. M., Fraser, C., Donnelly, C. A., Ghani, A. C. & Anderson, R. M. Public health risk from the avian H5N1 influenza epidemic. *Science* **304**, 968–969 (2004).
30. Farrington, C. P., Kanaan, M. N. & Gay, N. J. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics* **4**, 279–295 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful for comments and data suggestions from B. Bolker, J. Edmunds, N. Ferguson, A. Galvani, R. Gani, N. Gay, J. Gog, B. Grenfell, H. Hethcote, D. Heymann, A. Hubbard, N. Jewell, J. Lauer, R. May, T. Porco, C. Roth, D. Smith and B. Williams. We thank R. Gani for providing unpublished data from a previous publication, and L. Matthews for sharing work ahead of print. Our research was supported by the NSF, NIH-NIDA, the James S. McDonnell Foundation, the NSF/NIH Ecology of Infectious Disease Program, and the South African Centre for Epidemiological Modelling and Analysis (SACEMA).

Author Contributions J.O.L.-S. and W.M.G. conceived the study. J.O.L.-S. collected and analysed outbreak data, conducted dynamic modelling, and drafted and revised the text. S.J.S. conducted formal analysis of branching processes and control measures. W.M.G. provided technical input on superspreading and control analyses. All authors contributed conceptually, and edited or commented on the text.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.O.L.-S. (jls@nature.berkeley.edu).

LETTERS

Specification of astrocytes by bHLH protein SCL in a restricted region of the neural tube

Yuko Muroyama^{1†}, Yuko Fujiwara², Stuart H. Orkin^{1,2} & David H. Rowitch^{1,3}

Astrocytes are the most abundant and functionally diverse glial population in the vertebrate central nervous system (CNS)¹. However, the mechanisms underlying astrocyte specification are poorly understood. It is well established that cellular diversification of neurons in the embryo is generated by position-dependent extrinsic signals and combinatorial interactions of transcription factors that direct specific cell fates by suppressing alternative fates². It is unknown whether a comparable process determines embryonic astrocyte identity. Indeed, astrocyte development is generally thought to take place in a position-independent manner^{3,4}. Here we show multiple functions of *Stem cell leukaemia* (*Scl*, also known as *Tal1*), which encodes a basic helix–loop–helix (bHLH) transcription factor, in the regulation of both astrocyte versus oligodendrocyte cell fate acquisition and V2b versus V2a interneuron cell fate acquisition in the p2 domain of the developing vertebrate spinal cord. Our findings demonstrate a regionally restricted transcriptional programme necessary for astrocyte and V2b interneuron development, with striking parallels to the involvement of SCL in haematopoiesis. They further indicate that acquisition of embryonic glial subtype identity might be regulated by genetic interactions between SCL and the transcription factor *Olig2* in the ventral neural tube.

Astrocytes have traditionally been viewed as a neural cell population with structural and supportive roles such as the supply of essential substrates to neurons and participation in the blood–brain barrier¹. It is now recognized that certain astrocytic cells have stem cell properties^{4,5}, whereas others directly modulate the development and activity of the synapse⁶. Despite an enhanced understanding of the functional diversity of astroglia, it is generally believed that they develop in a uniform manner. For example, most studies of astrocyte differentiation have concentrated on the involvement of transcriptional mechanisms and signalling pathways with common roles throughout the CNS^{7–9}. The findings of enhanced astrocyte generation by negative-HLH proteins⁷ or in the absence of proneural bHLH function¹⁰, in particular, have led to the proposal that astrocytes develop ‘by default’^{3,4}.

In contrast, specification of embryonic oligodendrocyte precursors (OLPs) has much in common with motor neurons. During CNS development, extrinsic signals from organizing centres, such as the floorplate and roofplate, instruct naive progenitor cells in the ventricular zone to adopt cellular identities unique to their position in the embryo² (Fig. 1a). An early indication of pattern formation in the neural tube is the restricted expression of homeodomain and bHLH proteins in discrete domains along the dorsal–ventral axis.

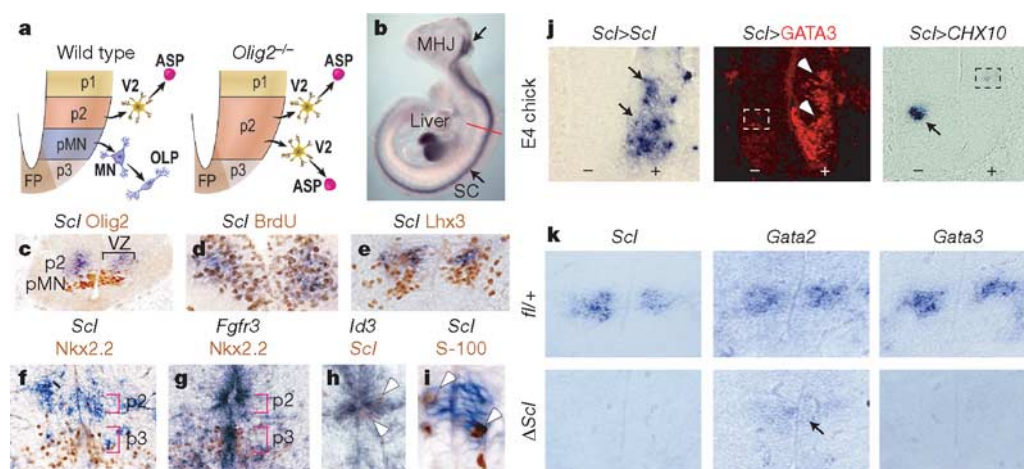


Figure 1 | *Scl* expression in p2 precursors and regulation of V2b interneuron development. **a**, Summary of p2 and pMN neuron and glial subtype production in wild-type and *Olig2*^{-/-} embryos. FP, floorplate. **b**, Whole-mount ISH shows *Scl* expression pattern in 10.5 d.p.c. mouse embryo. The red bar indicates forelimb level, the focus of subsequent analysis. SC, spinal cord; MHJ, mid-hindbrain junction. **c–e**, Expression of *Scl* (ISH) compared with immunohistochemistry (IHC) for *Olig2* (**c**), *BrdU* (**d**) and *Lhx3* (**e**) at 10.5 d.p.c. **f–i**, At 14.5 d.p.c., *Scl* expression (**f**) remains localized to p2

(*Nkx2.2* demarcates p3), coinciding with *Fgfr3* (**g**), *Id3* (**h**) and S-100 (**i**). **j**, Forced expression of *Scl* in E2 chick causes induction of GATA3⁺ V2b interneurons (arrows) and repression of CHX10⁺ V2a interneurons at E4 (representative of more than five embryos). **k**, SCL function is necessary for V2b interneuron development as shown by attenuation of *Gata2* (arrow) and loss of *Gata3* expression at 11.5 d.p.c. in ΔScl mutants ($n = 2$) compared with controls ($n = 2$).

¹Department of Pediatric Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. ²Howard Hughes Medical Institute and Divisions of Hematology and ³Division of Newborn Medicine, Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115, USA. [†]Present address: Department of Developmental Biology, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8670, Japan.

One such domain, the 'pMN', gives rise to motor neurons and OLPs in successive waves under the control of Sonic hedgehog (Shh) signalling^{11,12} (Fig. 1a). Expression of *Olig2* is specific to the pMN domain until a gestational age of 12.5 days post coitum (d.p.c.) in the mouse, and *Olig2* function is essential for the development of motor neurons and OLPs^{13–15}. In the absence of *Olig* function, progenitor cells that occupy the pMN location instead adopt p2 identity and produce V2 interneurons and astrocyte precursors (ASPs)^{14,15}. Understanding the molecular basis of this binary cell fate choice (oligodendrocyte versus astrocyte) in the neural tube would address fundamental questions in glial biology. It is possible that *Olig2* represses astrocyte fate by directing an alternative fate. In contrast, the finding of a localized astrogenic transcriptional programme in p2 would contravene current assumptions that astrocyte development occurs by default and that it is homogeneous throughout the CNS.

V2 interneurons, which establish communication between motor neurons in different spinal cord segments, are further divided into V2a (Chx10-positive or Chx10⁺) and V2b (GATA2/3-positive or GATA2/3⁺) groups^{16–18}. However, regulation of the production of V2 interneurons subtypes is as yet incompletely understood. *Scl*, which encodes a bHLH protein essential for haematopoietic stem cell specification and differentiation of the red-cell and megakaryocytic lineages^{19,20}, is expressed in close association with GATA2 and GATA3 in V2b interneurons^{16,17,21}. At 10.5 d.p.c., *Scl* is expressed principally in haematopoietic progenitors of the liver and throughout the anterior–posterior axis of developing spinal cord and hindbrain (Fig. 1b). Expression starts at 10.5 d.p.c. in the spinal cord and is specific to the p2 domain²¹, dorsal to the *Olig2*-expressing pMN (Fig. 1c). *Scl* messenger RNA transcripts *in situ* were observed predominantly in cycling neural precursors in the lateral aspect of the ventricular zone (VZ) (Fig. 1d), and partly overlapped *Lhx3*, a marker of V2a/b interneurons, in p2 (Fig. 1e). Glial acidic fibrillary protein (GFAP) is inappropriate for an analysis of astrogenesis in the murine neural tube because its expression does not begin until about 15.5–18.5 d.p.c.. Early astrocyte populations have been identified by the expression of *Fgfr3* (ref. 22) and *Id3* (ref. 23); moreover, S-100 has been shown to mark ASPs localized to the p2 region of the 14.5-d.p.c. mouse spinal cord²⁴. At 14.5 d.p.c., *Scl* VZ expression was detected in p2 (Fig. 1f) and coincided with *Fgfr3*, *Id3* and S-100 (Fig. 1g–i). These data indicate *Scl* expression in precursors for V2 interneurons and a subset of VZ cells that could represent immature astrocytes.

To test whether SCL regulates the fate of V2 interneurons, we overexpressed full-length mouse *Scl* in embryonic day (E)2 chicken spinal cord *in ovo*. We observed ectopic production of GATA3⁺ V2b interneurons and suppression of endogenous CHX10⁺ V2a interneurons on the electroporated side (Fig. 1j). To study requirements for *Scl* function, we used tissue-specific ablation because *Scl*^{−/−} animals die in the embryonic period¹⁹. Mice carrying a floxed *Scl*²⁵ were intercrossed with the *NesCre* line, which carries bacteriophage P1 *cre* recombinase under the control of nestin regulatory sequences²⁶, for the purpose of generating *NesCre* × *Scl*^{fl/fl} progeny (hereafter called ΔScl). From 10 d.p.c. onwards, we observed efficient ablation of *Scl* expression at forelimb levels of the spinal cord (Fig. 1k). As a consequence, GATA3⁺ cells failed to develop and we observed that *Gata2* levels were nearly undetectable. Thus, *Scl* function is necessary and sufficient for V2b interneuron development and is required for the maintenance of normal levels of *Gata2*. Because *Nestin-cre* expression can be mosaic²⁴, our data do not rule out an absolute requirement for *Scl* in the regulation of *Gata2*.

After V2 interneuron production, p2 progenitors give rise to a wave of ASPs^{15,24}. We observed a significant decrease in p2-associated S-100/S-100 β -expressing cells in ΔScl mice at 14.5 d.p.c. compared with controls (Fig. 2a, Supplementary Fig. 1a). Rare *Scl*/S-100⁺ cells were detected, indicating that some cells might escape *NesCre* excision (Supplementary Fig. 1b). In addition, a marked decrease in expression levels of *Fgfr3* and *Id3* at 14.5 d.p.c. (Fig. 2b, c) and 16.5 d.p.c. (Supplementary Fig. 1c) was observed. We next determined whether

decreases in ASP in ΔScl mice were coupled to increased production of OLPs. At 14.5 d.p.c., OLPs typically have migrated from the ventral VZ, but only sparsely populate dorsal regions of wild-type spinal cord (Fig. 2d, e). In contrast, we observed increased numbers of *Olig2*⁺ cells that were dorsally located in 14.5 d.p.c. ΔScl mice (Fig. 2d). Similar results were obtained with the OLP marker *Pdgfra* (Fig. 2e).

To determine whether SCL function is sufficient for specification of astroglia we used forced expression of mouse *Scl* in the chick neural tube (Fig. 2f). As shown in Fig. 2g, we observed repression of endogenous *Olig2* and ectopic development of S-100-expressing cells at the level of the pMN domain itself, and that *Scl* induced *FGFR3* and *ID3* (Fig. 2h, i). Astrocyte-specific expression of the clusterin, cystatin3 and brain glycogen phosphorylase genes has been reported in the mouse brain²⁷, and forced expression of *Scl* resulted in robust induction of these markers (Supplementary Fig. 1d) and

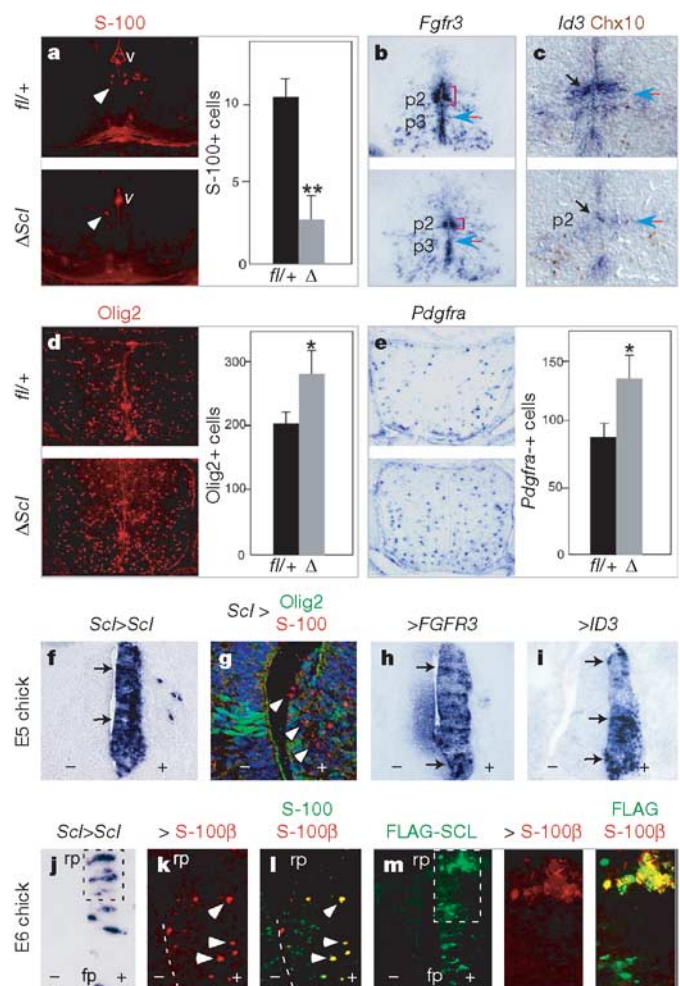


Figure 2 | SCL is necessary and sufficient for ASP development and represses OLP. **a**, A significant (two asterisks, $P < 0.01$) decrease in p2-associated S-100⁺ cells (white arrows) was observed in ΔScl mice ($n = 3$) compared with controls ($fl/+$ or $+/+$, $n = 4$). **b**, **c**, *Fgfr3* (dorsal limit of p3 indicated) and *Id3* levels (p2 region indicated by Chx10⁺ cells) (**c**) were decreased. **d**, **e**, Numbers of OLPs (*Olig2*⁺ or *Pdgfra*⁺) were significantly (asterisk, $P < 0.001$) increased in ΔScl mice. Histograms show cell numbers per histological section (means $\pm$ s.e.m.). **f–m**, Forced expression of *Scl* *in ovo* (**f**) leads to repression of *Olig2* and induction of S-100⁺ cells (arrows) in the former pMN (**g**) and upregulation of *FGFR3* (**h**) and *ID3* (**i**) at E5. **j–m**, SCL induces ASP (S-100 β /S-100⁺) in the dorsal neural tube at E6. Forced expression of *Scl* or *Scl*-FLAG was directed to the dorsal neural tube, and areas adjacent to the roof plate (rp, boxed in **j**, **m**) were analysed with the indicated markers (**k–m**).

S-100/S-100 β -expressing cells (Fig. 2j–m) throughout the chick neural tube, including dorsal progenitors. Taken together, our findings show that SCL function is necessary and sufficient for p2-associated astrocyte development and that it represses the production of OLPs.

These results raised the possibility of interactions between SCL and Olig2 in the ventral neural tube. Loss of V2b interneurons in ΔScl mutants was not associated with cell death (data not shown) and the number of V2a interneurons was significantly increased (Fig. 3a), indicating a possible fate switch (see Supplementary Fig. 2). Strikingly, we observed a dorsal expansion of Olig2 expression

(Fig. 3b) and increased numbers of Hb9⁺ MNs in ΔScl mutants (Fig. 3b). Conversely, forced expression of SCL in the chick neural tube repressed Olig2 (Fig. 3c) and the development of motor neurons; moreover, such effects were pMN-specific because the p3 marker Nkx2.2 was unaffected. Electroporation of Olig2 repressed Scl (Fig. 3d), and ectopic ventral cells expressing Scl, Gata2 and Gata3 were observed in Olig2^{-/-} embryos at 10.5 d.p.c. (Fig. 3e). We conclude that SCL and Olig2 participate in cross-antagonistic interactions that regulate cell fate specification.

The requirement for SCL in both V2b/astrocyte specification and haematopoietic development raises the possibility of shared transcriptional pathways. Within the haematopoietic system, SCL functions in heterodimeric association with E proteins (such as E47) and acts with GATA proteins, LMO2 and its partner NLI (Idb), either with or without direct DNA binding through its basic region^{28,29}. Expression of a tethered SCL–E47 heterodimer recapitulated the full activity of SCL alone, including Olig2 repression, the induction of astrocytes (Fig. 3f) and V2b interneurons (not shown). This activity was dependent on DNA binding by SCL (Fig. 3f). As LMO4 and LMO2 (and GATA2/3 and NLI) are coexpressed with SCL in the pMN/p2 region^{17,18,21} (Y.M. and D.H.R., unpublished observations), we tested the potential role of SCL–LMO interactions within a nucleoprotein complex by expressing a mutant of SCL (F238–G) that fails to recruit LMO2 (refs 28, 29). This mutant is inactive in the chick neural tube assay (Fig. 3f), providing strong evidence that SCL functions in this context in a complex with E proteins and an LMO protein (either LMO2 or LMO4). Analysis of SCL fusions to VP16 or the engrailed repressor domain indicates that SCL acts as a transcriptional activator in the ventral neural tube (Supplementary Fig. 3). These findings indicate that SCL-mediated antagonism of Olig2 expression might occur at the transcriptional level and is most probably indirect (see Supplementary Fig. 3). Further, they reveal striking parallels between SCL regulatory interactions in astrocyte and haematopoietic development.

Our findings demonstrate critical roles for SCL in the development of astrocytes that are localized to the p2 domain of the neural tube and V2b interneurons. SCL both regulates the fidelity of late p2 progenitors and links V2b interneuron/astrocyte progeny production by a common transcriptional mechanism, which is dependent on DNA binding and requires interactions with LMO proteins, as previously shown for late SCL functions in the development of erythrocytes and megakaryocytes^{28,29}. Our results provide evidence that SCL engages in cross-antagonistic interactions with Olig2 to regulate cell fate, and indicate that bHLH ‘code’ might couple the acquisition of both glial and neuronal subtype identity in the ventral neural tube (Fig. 4). With regard to gliogenesis, such a mechanism is necessary to regulate the size of initial OLP and ASP pools and, by extension, the timing and scope of early neuron–glial interactions. In

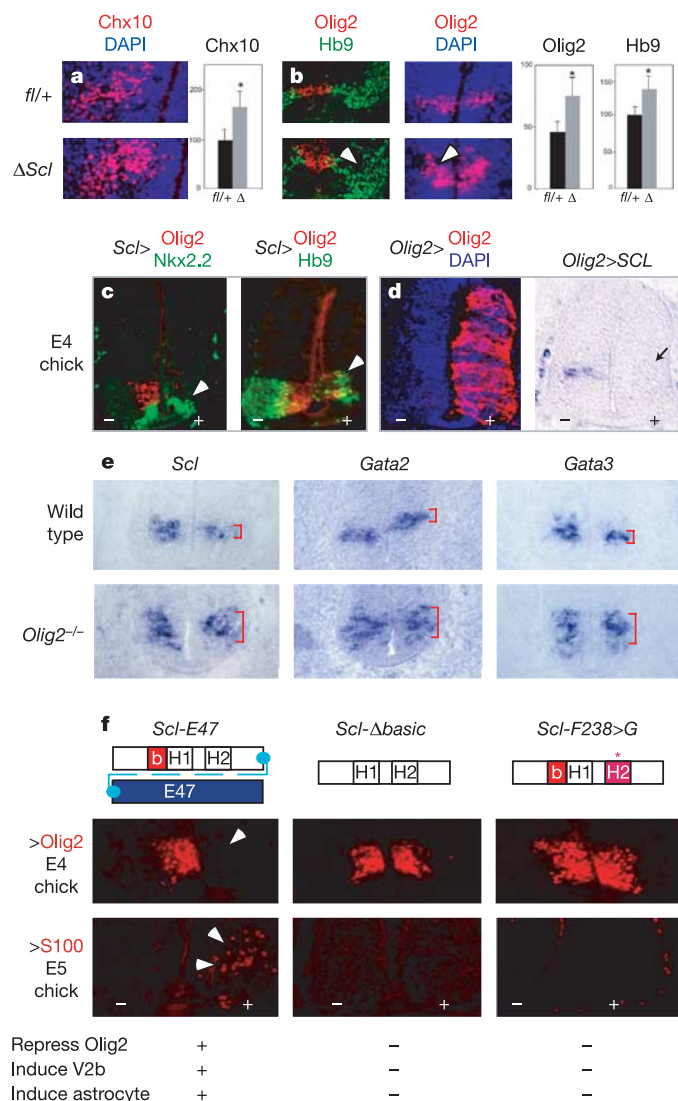


Figure 3 | Cross-antagonistic interactions between SCL and Olig2.

a, b. Significantly (asterisk, $P < 0.001$) increased numbers of Chx10⁺ (**a**), Olig2⁺ and Hb9⁺ (**b**) cells were observed in ΔScl mice ($n = 3$) compared with controls ($fl/+$ or $+/+$; $n = 4$) at 11.5 d.p.c. In **b**, note ectopic dorsal Olig2⁺ cells in lateral VZ (arrowhead) where Scl is normally expressed (Fig. 1c). Histograms show cell numbers per histological section (mean $\pm$ s.e.m.). **c.** SCL represses Olig2 expression and motor neurons but not Nkx2.2 expression. **d.** Olig2 represses SCL (arrow). **e.** Ventral expansion of Scl and Gata2/3 expression in Olig2^{-/-} 10.5 d.p.c. mouse. **f.** SCL–E47 tethered heterodimer recapitulates SCL activity including Olig2 repression (white arrowheads, top panels) and S-100 induction (bottom panels). Despite robust expression (not shown), Scl mutants that lack the basic region required for DNA binding²⁸ or a (Phe 238) residue for LMO interactions²⁹ are inactive. Activities are summarized at bottom: +, present; –, absent.

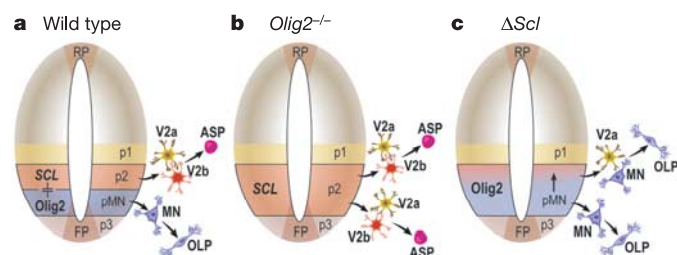


Figure 4 | SCL and Olig2 regulate neuronal and glial subtype development in the ventral neural tube. **a.** SCL cross-antagonistic interactions with Olig2 maintain p2 precursors that give rise to V2b interneurons and ASP. The switch from production of neurons to glia in p2 requires function of Sox9 (ref. 24) and is complex³⁰. **b.** Loss of Olig2 function results in ventral expansion of Scl expression, V2a/b interneurons and ASP¹⁵. **c.** In ΔScl mutants, dysregulated p2 precursors produce V2a interneurons and motor neurons but not V2b interneurons.

the adult mouse brain, *Scl* is expressed in the subventricular zone, a source of persistent neural progenitors, and the corpus callosum (Y.M. and D.H.R., unpublished observations), which is consistent with additional roles for SCL during astrogenesis in late-developing neural populations. Although SCL overexpression is sufficient to promote astrocyte development throughout the embryonic neural tube, it differs from transcription factors with pan-glial roles in the embryo (for example, Sox9 (ref. 24)) because it is normally expressed in a highly restricted CNS domain, most probably regulated by Shh and other organizing signals. Thus, our findings identify a position-restricted astrogenic transcriptional mechanism in the embryo. It is possible that the heterogeneous origins of precursors might contribute to the ultimate functional and molecular diversity of astrocytes in the vertebrate CNS.

METHODS

Transgenic mice and neural-tube electroporation *in ovo*. Mice with targeted mutation of *Scl* and *Olig2* loci have been described^{14,25}. Because the NesCre line²⁶ is subject to mosaic expression, it was important to use NesCre × *Scl*^{fl/+} males mated to *Scl*^{fl/fl} females to achieve ablation in NesCre × *Scl*^{fl/fl} progeny²⁴. Electroporation in E2 chick embryos was performed as described in Supplementary Methods, and the results shown are representative of more than five independent experiments. In each case, ectopic expression was confirmed in adjacent sections by *in situ* hybridization (ISH) or immunohistochemistry (IHC).

Tissue preparation, analysis and imaging. Embryos were fixed in 4% para-formaldehyde in PBS pH 7.0, embedded in optimal cutting temperature (OCT) compound (Sakura Finetek) and cryosectioned at 10–14 µm. Antibodies and IHC with antigen retrieval and probes used in ISH are described in Supplementary Methods. In some cases, ISH was followed by IHC for assessing *Scl* expression with respect to other markers. Detailed instructions are available from the authors on request. Standard photomicrographic images were collected on a Nikon E600 microscope and Nikon or SPOT digital cameras. Confocal micrographs were collected on a Zeiss LSM510 microscope.

Cell counts and statistical measures. Histograms show cell numbers per histological section (means ± s.e.m.) derived from counts from six anterior–posterior spinal cord levels from at least two embryos of any given genotype (for details of genotype and *n*, see the text). The level of significance was determined by Student's *t*-test.

Received 20 July; accepted 19 August 2005.

- Kettenmann, H. & Ransom, B. R. Electrical coupling between astrocytes and between oligodendrocytes studied in mammalian cell cultures. *Glia* **1**, 64–73 (1988).
- Jessell, T. M. Neuronal specification in the spinal cord: inductive signals and transcriptional codes. *Nature Rev. Genet.* **1**, 20–29 (2000).
- Ross, S. E., Greenberg, M. E. & Stiles, C. D. Basic helix-loop-helix factors in cortical development. *Neuron* **39**, 13–25 (2003).
- Doetsch, F. The glial identity of neural stem cells. *Nature Neurosci.* **6**, 1127–1134 (2003).
- Alvarez-Buylla, A., Garcia-Verdugo, J. M. & Tramontin, A. D. A unified hypothesis on the lineage of neural stem cells. *Nature Rev. Neurosci.* **2**, 287–293 (2001).
- Christopherson, K. S. *et al.* Thrombospondins are astrocyte-secreted proteins that promote CNS synaptogenesis. *Cell* **120**, 421–433 (2005).
- Nakashima, K. *et al.* BMP2-mediated alteration in the developmental pathway of fetal mouse brain cells from neurogenesis to astrocytogenesis. *Proc. Natl Acad. Sci. USA* **98**, 5868–5873 (2001).
- Hermanson, O., Jepsen, K. & Rosenfeld, M. G. N-CoR controls differentiation of neural stem cells into astrocytes. *Nature* **419**, 934–939 (2002).
- Song, M. R. & Ghosh, A. FGF2-induced chromatin remodeling regulates CNTF-mediated gene expression and astrocyte differentiation. *Nature Neurosci.* **7**, 229–235 (2004).
- Nieto, M., Schuurmans, C., Britz, O. & Guillemot, F. Neural bHLH genes control the neuronal versus glial fate decision in cortical progenitors. *Neuron* **29**, 401–413 (2001).
- Ericson, J., Morton, S., Kawakami, A., Roelink, H. & Jessell, T. M. Two critical periods of Sonic Hedgehog signalling required for the specification of motor neuron identity. *Cell* **87**, 661–673 (1996).
- Orentas, D. M., Hayes, J. E., Dyer, K. L. & Miller, R. H. Sonic hedgehog signalling is required during the appearance of spinal cord oligodendrocyte precursors. *Development* **126**, 2419–2429 (1999).
- Novitsch, B. G., Chen, A. I. & Jessell, T. M. Coordinate regulation of motor neuron subtype identity and pan-neuronal properties by the bHLH repressor Olig2. *Neuron* **31**, 773–789 (2001).
- Lu, Q. R. *et al.* Common developmental requirement for Olig function indicates a motor neuron/oligodendrocyte connection. *Cell* **109**, 75–86 (2002).
- Zhou, Q. & Anderson, D. J. The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification. *Cell* **109**, 61–73 (2002).
- Zhou, Y., Yamamoto, M. & Engel, J. D. GATA2 is required for the generation of V2 interneurons. *Development* **127**, 3829–3838 (2000).
- Karunaratne, A., Hargrave, M., Poh, A. & Yamada, T. GATA proteins identify a novel ventral interneuron subclass in the developing chick spinal cord. *Dev. Biol.* **249**, 30–43 (2002).
- Thaler, J. P., Lee, S. K., Jurata, L. W., Gill, G. N. & Pfaff, S. L. LIM factor Lhx3 contributes to the specification of motor neuron and interneuron identity through cell-type-specific protein-protein interactions. *Cell* **110**, 237–249 (2002).
- Porcher, C. *et al.* The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86**, 47–57 (1996).
- Gering, M., Rodaway, A. R., Gottgens, B., Patient, R. K. & Green, A. R. The SCL gene specifies haemangioblast development from early mesoderm. *EMBO J.* **17**, 4029–4045 (1998).
- Smith, E., Hargrave, M., Yamada, T., Begley, C. G. & Little, M. H. Coexpression of SCL and GATA3 in the V2 interneurons of the developing mouse spinal cord. *Dev. Dyn.* **224**, 231–237 (2002).
- Pringle, N. P. *et al.* *Fgfr3* expression by astrocytes and their precursors: evidence that astrocytes and oligodendrocytes originate in distinct neuroepithelial domains. *Development* **130**, 93–102 (2003).
- Tzeng, S. F. & de Vellis, J. Id1, Id2, and Id3 gene expression in neural cells during development. *Glia* **24**, 372–381 (1998).
- Stolt, C. C. *et al.* The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes Dev.* **17**, 1677–1689 (2003).
- Mikkola, H. K. *et al.* Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia SCL/tal-1 gene. *Nature* **421**, 547–551 (2003).
- Tronche, F. *et al.* Disruption of the glucocorticoid receptor gene in the nervous system results in reduced anxiety. *Nature Genet.* **23**, 99–103 (1999).
- Bachoo, R. M. *et al.* Molecular diversity of astrocytes with implications for neurological disorders. *Proc. Natl Acad. Sci. USA* **101**, 8384–8389 (2004).
- Porcher, C., Liao, E. C., Fujiwara, Y., Zon, L. I. & Orkin, S. H. Specification of hematopoietic and vascular development by the bHLH transcription factor SCL without direct DNA binding. *Development* **126**, 4603–4615 (1999).
- Schlaeger, T. M. *et al.* Decoding hematopoietic specificity in the helix-loop-helix domain of the transcription factor SCL/Tal-1. *Mol. Cell. Biol.* **24**, 7491–7502 (2004).
- Rowitch, D. H. Glial specification in the vertebrate neural tube. *Nature Rev. Neurosci.* **5**, 409–419 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Ligon, T. Schlaeger, M. Goulding and Q. Ma for technical suggestions and for reagents and comments. S.H.O. is an Investigator of the Howard Hughes Medical Institute. This work was supported in part by a Center of Excellence Award from NIDDK (to S.H.O.) and grants (to D.H.R.) from NINDS, the NCI and the National Multiple Sclerosis Society.

Author Contributions Y.M. and Y.F. contributed equally to this work. Y.F. generated the transgenic mice and performed initial phenotypic analysis. Y.M. performed chick electroporation studies and all neural tube analyses. Y.M., S.H.O. and D.H.R. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.H.O. (stuart_orkin@dfci.harvard.edu) or D.H.R. (david_rowitch@dfci.harvard.edu).

LETTERS

Control of B-cell responses by Toll-like receptors

Chandrashekhar Pasare¹ & Ruslan Medzhitov¹

Toll-like receptors (TLRs) detect microbial infection and have an essential role in the induction of immune responses^{1–3}. TLRs can directly induce innate host defence responses, but the mechanisms of TLR-mediated control of adaptive immunity are not fully understood. Although TLR-induced dendritic cell maturation is required for activation of T-helper (T_H) cells⁴, the role of TLRs in B-cell activation and antibody production *in vivo* is not yet known. Here we show that activation and differentiation of T_H cells is not sufficient for the induction of T-dependent B-cell responses. We find that, in addition to CD4⁺ T-cell help, generation of T-dependent antigen-specific antibody responses requires activation of TLRs in B cells.

MyD88 (myeloid differentiation primary response gene 88) is a TLR signalling adaptor critical for TLR-dependent induction of adaptive immune responses^{5,6}. T_H cell activation and production of interferon (IFN)- γ are deficient in MyD88 knockout mice^{3,5,6}. Consistent with this defect in T_H cell activation, antigen-specific IgM and IgG1 antibody responses to T-dependent antigens are considerably reduced, whereas IgG2 antibody responses are completely abolished in MyD88 knockout mice (Fig. 1a and data not shown). The steady-state levels of total serum IgM, IgG1, IgG2c and IgG3 in naive MyD88 knockout mice are reduced compared to levels in wild-type mice (Supplementary Fig. 1), despite the fact that B-cell populations are normal and phenotypically comparable between these mice

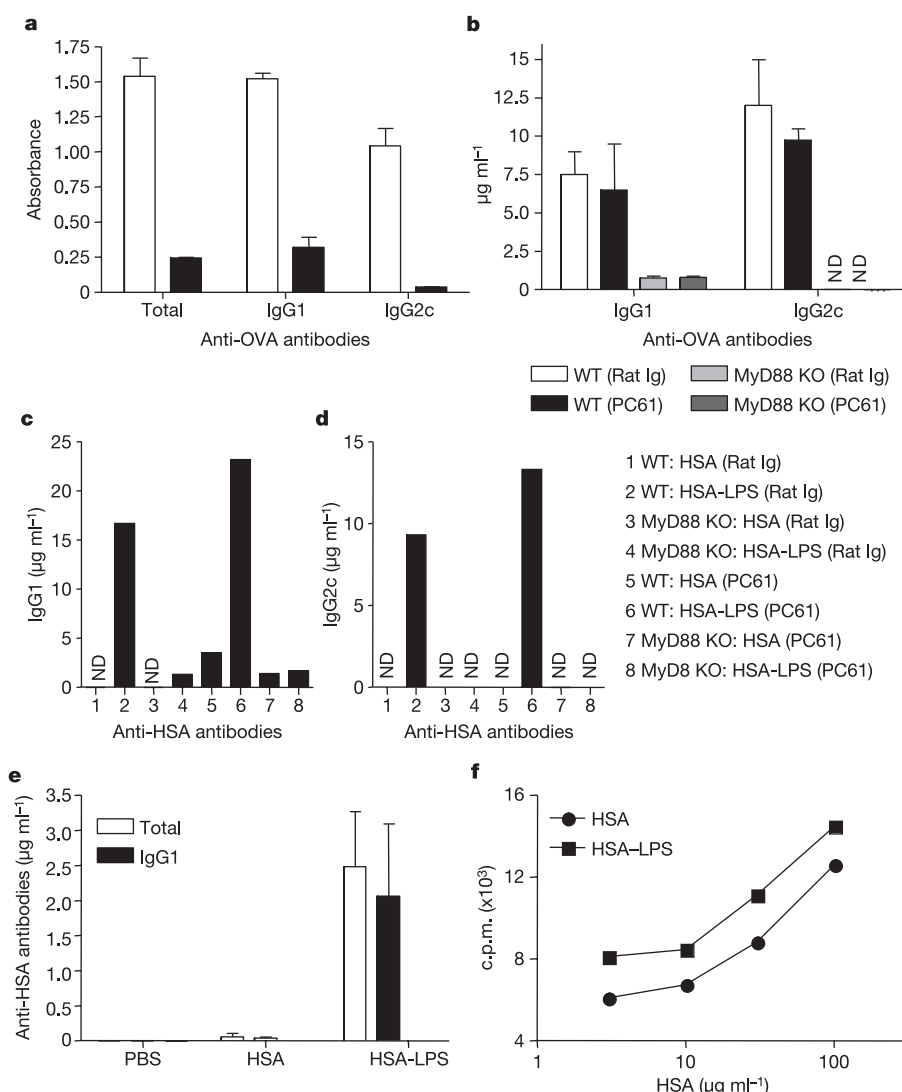


Figure 1 | T-cell activation is insufficient to induce normal antibody responses. **a**, Wild-type (white bars) and MyD88 knockout (black bars) mice (three per group) were immunized via the footpads with OVA-LPS in IFA and were bled 14 days later to assay for OVA-specific antibodies. Data are shown as absorbance values at 490 nm wavelength at appropriate dilutions and are representative of six independent experiments. **b**, Sera from control and T_R-depleted mice (four per group) immunized and bled as in **a** were analysed for OVA-specific antibodies. Data are representative of four independent experiments. KO, knockout. **c**, **d**, Control and T_R-depleted mice were immunized intraperitoneally with HSA or HSA-LPS adsorbed on alum and bled on day 12 for estimation of anti-HSA antibodies. Data are representative of two independent experiments. **e**, **f**, μMT mice were immunized using HSA-LPS in IFA (in the footpads) followed by B-cell (WT) transfer after 90 days. Mice were divided into three groups (four mice each) and received PBS, HSA or HSA-LPS adsorbed on alum via the intraperitoneal route. Anti-HSA antibody levels in sera after 12 days (**e**) and *in vitro* CD4⁺ T-cell activation after 14 days (**f**) are shown. Data are representative of two independent experiments. c.p.m., counts per min; ND, non-detectable. Error bars in **a**, **b**, **e** indicate standard error.

¹Howard Hughes Medical Institute and Section of Immunobiology, Yale University School of Medicine, 300 Cedar Street New Haven, Connecticut 06510, USA.

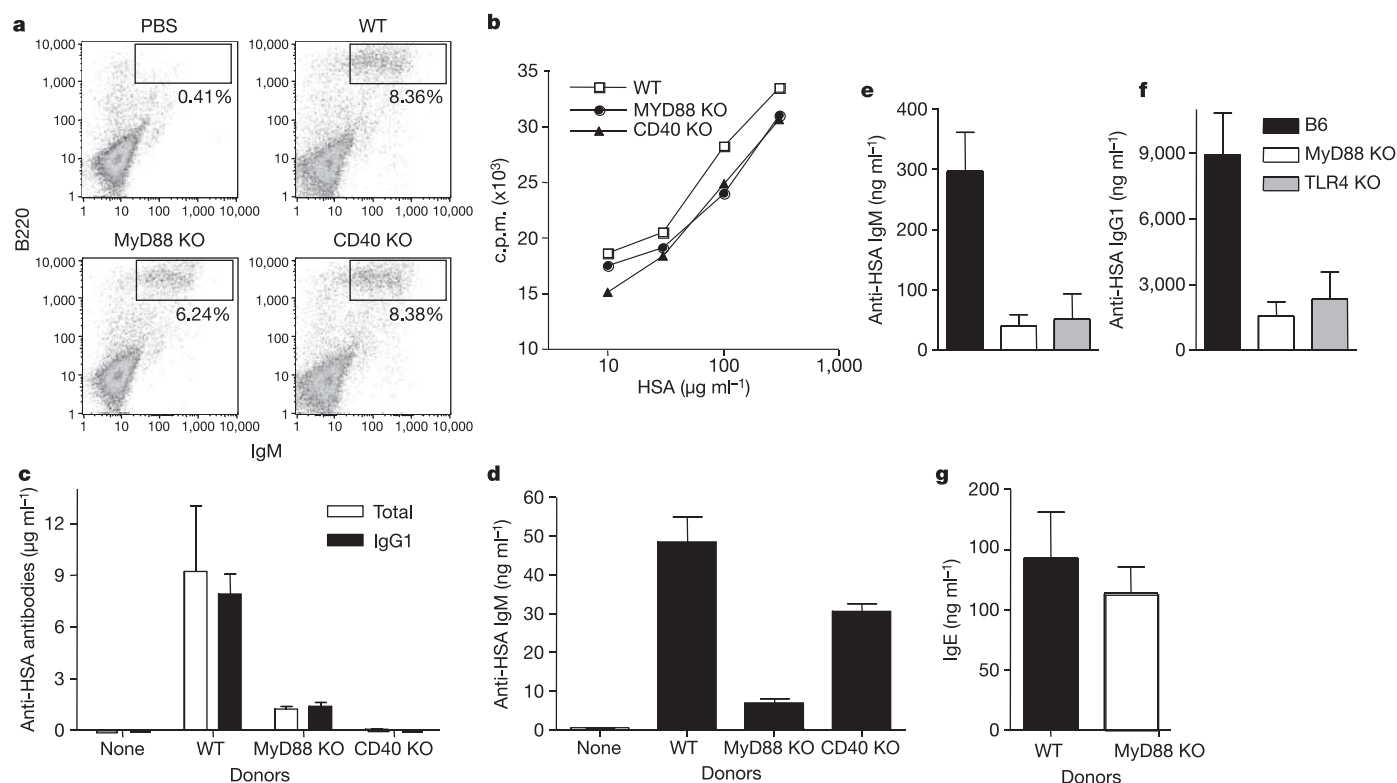


Figure 2 | Defective antibody production by MyD88 knockout B cells is B-cell intrinsic. **a**, One day after B-cell transfer, μ MT mice were bled and cells were stained for B220 and IgM. Mice were immunized 24 h after B-cell transfer with HSA-LPS on alum (intraperitoneal). **b**, After 12 days, CD4⁺ T cells were re-stimulated *in vitro* with titrating doses of HSA and irradiated B cells. **c**, **d**, Sera were analysed for anti-HSA antibodies. Data are representative of five independent experiments. **e**, μ MT mice (four per group) received B cells from the indicated donors followed by

intraperitoneal immunization of HSA-LPS on alum. Twelve days later sera were analysed for anti-HSA IgM (**e**) and IgG1 (**f**). Data are representative of two independent experiments. **g**, μ MT mice (three per group) received B cells from the indicated donors followed by intraperitoneal immunization of HSA-LPS on alum. Sera were analysed after 12 days for total anti-IgE antibodies. Control mice that received B cells but no immunization showed no detectable IgE in their sera. Error bars in **c–g** indicate standard error.

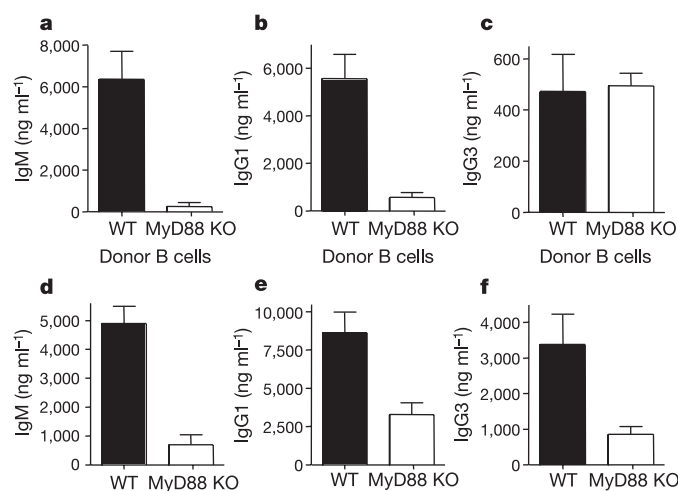


Figure 3 | Differential requirements for TLR activation on B cells by T-dependent and T-independent antigens. B cells (25 million per mouse) from wild type and MyD88 knockout mice were transferred to μ MT mice (four per group) and after 24 h the recipient mice received intraperitoneal immunization of flagellin on alum (50 μg per mouse). Mice were bled after 12 days and sera were analysed for anti-flagellin IgM (**a**), IgG1 (**b**) and IgG3 (**c**). Data are representative of two independent experiments. In an independent experiment wild type and MyD88 knockout mice were immunized and bled as described above to assay for anti-flagellin antibodies (**d**, **e**, **f**). Error bars indicate standard error.

(Supplementary Fig. 2). One reason for this B-cell response defect is the deficient T_H cell activation and consequent lack of T-cell help for B-cell activation and isotype switch. Depletion of regulatory T (T_R) cells in MyD88 knockout mice restores T_H cell activation and IFN-γ production to wild-type levels (ref. 7 and data not shown). However, immunization after T_R depletion does not result in recovery of T-dependent B-cell responses in these mice (Fig. 1b–d). This result suggests that T_H activation and IFN-γ production are not sufficient for T-dependent antibody responses.

We previously showed that TLR signalling is not required for activation of memory T_H cells⁷. This allowed us to create an experimental scenario where memory T_H cells would be activated in the presence of an antigen but no TLR ligands would be available to engage TLRs on B cells. μ MT mice that lack mature B cells⁸ were immunized with human serum albumin and lipopolysaccharide (HSA-LPS) to generate HSA-specific memory T_H cells. After 90 days, these mice received purified naive B cells from wild-type mice and were immunized with either HSA alone, or HSA together with LPS. Mice that received HSA alone made a deficient anti-HSA antibody response compared to mice that received HSA and LPS (Fig. 1e). Comparable memory T_H cell responses were seen between the two groups of mice (Fig. 1f), consistent with our previous report⁷. These results demonstrate that even memory T_H cells cannot support T-dependent antibody responses, and suggest a requirement for direct B-cell stimulation by TLR ligands.

To determine whether antibody responses required TLR signalling in B cells or in other cell types, we performed B-cell transfer experiments. Purified B cells from wild type, MyD88 knockout, TLR4 knockout and CD40 knockout mice were transferred into

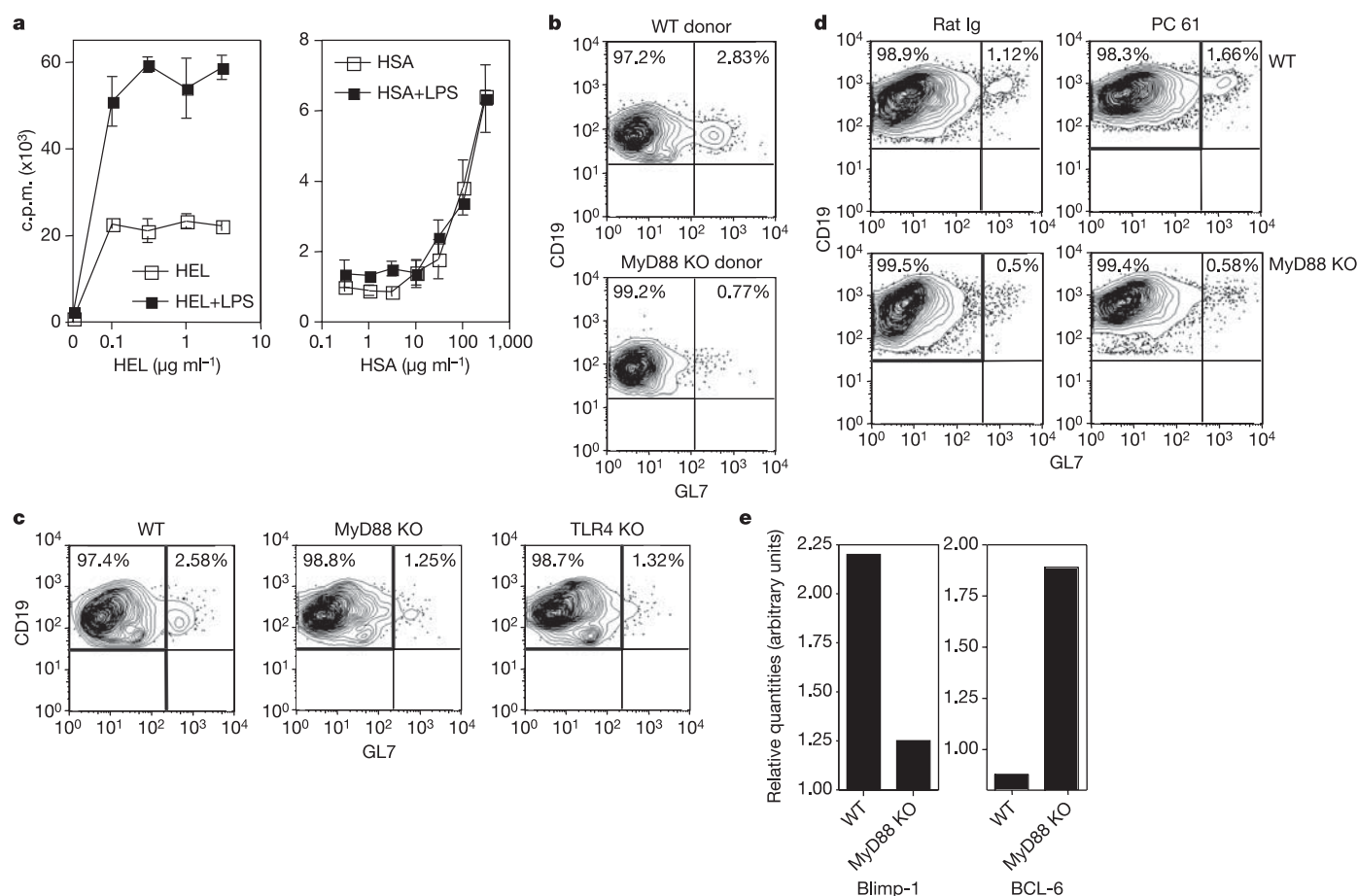


Figure 4 | TLR signalling influences several aspects of B-cell activation and differentiation. **a**, Anti-HEL immunoglobulin transgenic B cells, in the presence or absence of LPS (100 ng ml⁻¹), were incubated with titrating doses of HEL or HSA, and purified CD4⁺T cells from mice immunized with HEL-LPS or HSA-LPS emulsified in IFA, respectively. Cultures were pulsed with [³H]thymidine after 48 h and cells were harvested 12 h later to measure thymidine incorporation. Data represent two independent experiments. **b, c, d**, B cells (25 million) from the indicated donors were transferred into μ MT mice and the recipients were immunized a day later with HSA-LPS adsorbed on alum (intraperitoneal). Mice were killed 12 days later and cells from spleen were stained for CD19 and GL7. Populations of

CD19-expressing cells positive for GL7 are shown as contour plots. **d**, Control and T_R-depleted mice were immunized (intraperitoneal) with OVA-LPS on alum, and cells from spleen were stained as described above. Data in **b–d** are representative of three independent experiments. **e**, Wild type or T_R-depleted MyD88 knockout mice were immunized (intraperitoneal) with OVA-LPS on alum and GL7-positive cells were sorted after 12 days; cDNA was used for quantitative RT-PCR analysis. Relative quantities of Blimp-1 and BCL-6 messages (normalized to HPRT) in GL7-positive cells are shown. Data are representative of two independent experiments. Error bars in **a** show standard error.

μ MT mice that lack mature B cells⁸, followed by immunization with HSA and LPS. The B-cell transfers resulted in equivalent homing and survival of B cells in all mouse groups (Fig. 2a). The CD4⁺ T-cell responses against HSA (Fig. 2b) and IFN- γ and interleukin (IL)-13 production by activated T cells (Supplementary Fig. 3) were comparable in all the recipient μ MT mice. Analysis of antigen-specific antibody responses demonstrated that, although mice that received wild-type B cells were able to generate antigen-specific IgM and IgG responses, production of antibodies was considerably impaired in mice that received TLR4 knockout or MyD88 knockout B cells (Fig. 2c–f). Notably, induction of IgE response after HSA-LPS on alum immunization was comparable between mice that had received wild type or MyD88 knockout B cells (Fig. 2g). As expected, CD40 knockout B cells could only generate an IgM response, consistent with previous studies of CD40 knockout mice⁹ and patients with CD40 ligand deficiency¹⁰. The deficient IgM response by TLR4 knockout and MyD88 knockout B cells, but not by CD40 knockout B cells, indicates that TLRs and CD40 have distinct roles in B-cell activation and antibody class switch (Fig. 2d). TLR signalling in B cells is required for IgM and IgG production, whereas CD40 signalling is not required for IgM production, but is required for isotype

switch. Thus, at least for primary B-cell activation, TLRs and T-cell help have distinct but complementary roles.

We further investigated the requirement for TLR signalling in B cells for T-independent type II antibody responses. Flagellin can serve as both a T-dependent protein antigen and T-independent type II antigen (due to its ability to polymerize). In addition, flagellin is a ligand for TLR5 (ref. 11), which is expressed by B cells (Supplementary Fig. 4). Flagellin immunization of μ MT mice that had received wild-type B cells elicited strong anti-flagellin IgM, IgG1 (both T-dependent) and IgG3 (T-independent) responses (Fig. 3a–c). μ MT mice that had received MyD88 knockout B cells were deficient in anti-flagellin IgM and IgG1 responses (Fig. 3a, b). However, these mice mounted a normal IgG3 response comparable to mice transferred with wild-type B cells (Fig. 3c). Thus, TLR signalling in B cells is not required for induction of the IgG3 response. However, flagellin immunization of MyD88 knockout mice not only led to deficient IgM and IgG1 responses (Fig. 3d, e) but also revealed a deficient IgG3 response in comparison to wild-type mice (Fig. 3f), suggesting that TLR signalling in other cell types (possibly dendritic cells and macrophages), but not B cells, is required for generation of optimal IgG3 responses. This finding may explain, at least in part, the

requirement for accessory cells in the generation of T-independent type II antibody responses^{12,13}.

T-dependent B-cell response is a multi-step process^{14,15} and we investigated its regulation by TLR signalling in B cells. We first tested whether TLR signalling affected antigen processing and presentation by antigen-specific B cells. We used anti-hen egg lysozyme (HEL) B-cell receptor (BCR)-transgenic B cells¹⁶ to examine whether presentation of antigens taken up by receptor-mediated endocytosis would be affected by the presence of non-mitogenic levels of LPS. Although B cells are very efficient at presenting antigens taken up through the BCR, TLR4 ligation enhanced the presentation to HEL-specific T cells by several fold (Fig. 4a, left panel). Notably, LPS did not enhance presentation of antigen internalized by fluid phase endocytosis (Fig. 4a, right panel).

We next examined the role of TLR signalling in the differentiation of B cells into germinal centre cells. At day 12 after immunization splenic B cells in μ MT mice were analysed for expression of the germinal centre B-cell marker GL7 (ref. 17). Mice that received MyD88 knockout or TLR4 knockout B cells had a lower percentage of germinal centre B cells compared to mice that received wild-type B cells (Fig. 4b, c; see also Supplementary Fig. 5). Deficient B-cell activation and germinal centre formation were also seen in MyD88 knockout mice that were immunized after T_H depletion (Fig. 4d), even though T_H activation under these conditions is comparable to wild-type mice (ref. 7 and data not shown). BCL-6 (refs 18, 19) and Blimp-1 (refs 20, 21) are the master regulators of the germinal centre to plasma cell transition, and we tested whether TLR signalling in B cells regulates their expression. LPS stimulation of B cells *in vitro*

leads to strong induction of Blimp-1 (refs 22, 23 and data not shown). Estimation of relative levels of Blimp-1 and BCL-6 by quantitative polymerase chain reaction with reverse transcription (RT-PCR) revealed higher expression of Blimp-1 and lower expression of BCL-6 in GL7-positive cells in immunized wild-type mice when compared to GL7-positive cells from T_H -depleted and immunized MyD88 knockout mice (Fig. 4d). Taken together, these findings suggest that TLR signalling affects multiple stages of B-cell activation. Enhanced antigen presentation by antigen-specific B cells after TLR activation probably contributes to the development of the germinal centre reaction, and subsequent induction of Blimp-1 and concomitant suppression of BCL-6 (refs 24, 25) in germinal centre B cells by TLR ligands allows efficient differentiation into antibody-producing plasma cells.

Experiments described so far rely on B-cell transfers into lymphopenic hosts (μ MT mice). To ensure that the results and conclusions are not biased in some way by cell transfers, we developed a novel experimental system. We generated transgenic mice expressing MyD88 under the control of the CD11c promoter. These mice were then crossed to the MyD88 knockout background so that only dendritic cells, but not B cells, would be responsive to TLRs. Analysis of these mice revealed functional expression of MyD88 by dendritic cells but not by B cells (Supplementary Fig. 6). Immunization of CD11c-MyD88 transgenic mice with HSA and LPS induced T_H1 responses (Fig. 5a, b), but antigen-specific antibody production was deficient in these mice (Fig. 5c), demonstrating that TLR signalling in B cells is required for T-dependent antibody responses.

The results presented here reveal an unexpected requirement for TLR signalling in B cells for optimal antibody responses to T-dependent antigens. TLRs are thus involved in the control of antibody responses to T-dependent antigens in at least two ways. First, TLR signals induce dendritic cell maturation and T_H cell activation, which are required to provide T-cell help to B cells, primarily through CD40–CD40L interaction and cytokine secretion. Second, TLRs expressed on B cells have a direct role in B-cell activation and antibody production. This function of TLRs in B cells may help to determine the microbial origin of antigens recognized by the BCR and help direct the response against infectious agents. In addition, TLR signalling in memory B cells contributes to maintenance of serological memory²⁶. Notably, while TLR signalling is important for the induction of IgM, IgG1 and IgG2c responses (Figs 1 and 2), it is dispensable for the induction of IgE responses (Fig. 2g)⁶. T_H2 and IgE responses are initiated by and are protective against multicellular parasites, and their recognition does not seem to be mediated by TLRs²⁷. Normal levels of IgA in the serum of MyD88 knockout mice suggest that TLR signalling might be dispensable for induction of IgA as well (Supplementary Fig. 1).

As with other pathways of lymphocyte activation, inappropriate engagement of TLRs on B cells can lead to autoimmune responses under certain conditions²⁸. Accordingly, engagement of TLRs on self-reactive B cells fails to promote B-cell activation due to a tolerogenic signal transduced by the self reactive BCR²⁹. On the other hand, coupled BCR–TLR recognition of microbial antigens is required for optimal antibody responses³⁰.

METHODS

Mice. MyD88 knockout mice and TLR4 knockout mice on a C57BL/6 background were bred and maintained at the animal facility of Yale University School of Medicine. B-cell-deficient (μ MT), CD40 knockout and C57BL/6 mice were obtained from Jackson laboratories. Transgenic mice expressing the MyD88 gene under the control of the CD11c promoter (CD11c-MyD88 Tg) were generated at Yale Animal Genomics services. The CD11c-MyD88 transgenic mice were backcrossed to MyD88 knockout mice. All mice were used at 8–12 weeks of age. **Reagents and antibodies.** LPS, ovalbumin (OVA) and endotoxin-free HSA were purchased from Sigma. Anti-CD8 (TIB 105), anti-CD4 (GK1.5), anti-MAC-1 (TIB-128) and anti-Thy1 (Y19) monoclonal antibodies were used as hybridoma supernatants (ATCC). Fluorescein isothiocyanate (FITC) anti-CD4, phycoerythrin (PE)-anti-CD25, PE-anti B220, FITC anti-IgM and purified anti-CD25

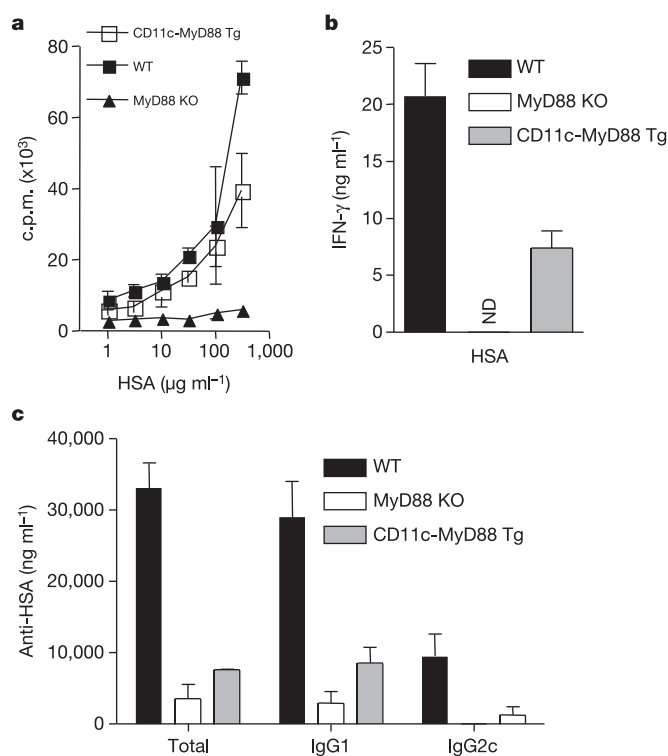


Figure 5 | Activation of dendritic cells and T cells is insufficient to induce normal antibody responses in a MyD88 complementation model. a, Mice were immunized via the footpads using HSA-LPS emulsified in IFA, and 10 days later purified CD4⁺T cells from draining lymph nodes were incubated with irradiated wild-type B cells and titrating doses of HSA to assay for T-cell proliferation. **b**, Supernatants from cultures were collected after 72 h to measure IFN- γ secretion by activated T cells. **c**, Sera from immunized mice (three per group) were analysed for anti-HSA antibodies. Data are representative of two independent experiments. Error bars show standard error. ND, non-detectable.

(clone PC61) were all purchased from BD Biosciences. Anti-CD4 and anti-CD19 microbeads were purchased from Miltenyi Biotec. Polyclonal anti-mouse immunoglobulin, anti-mouse immunoglobulin biotin, anti-mouse IgM biotin, anti-mouse IgG1 biotin and anti-mouse IgG2c biotin were purchased from Southern Biotechnology Associates. Mouse cells were cultured in complete RPMI-1640 supplemented with 10% FCS, 100 U ml⁻¹ penicillin, 100 µg ml⁻¹ streptomycin, 2 mM L-glutamine, 10 mM HEPES and 1 mM sodium pyruvate (all from Sigma).

In vivo depletion of T_R cells. *In vivo* depletion of CD4⁺CD25⁺ T cells was achieved by intravenous injection of 100 µg of monoclonal anti-CD25 antibody (clone PC61).

Immunization. Mice were immunized in hind footpads with 50 µg per mouse of OVA (Sigma) or HSA and 5 µg per mouse of LPS emulsified in incomplete Freund's adjuvant (IFA) (Sigma). For certain experiments mice were immunized intraperitoneally with HSA (50 µg) or HSA (50 µg) and LPS (5 µg) co-adsorbed on alum. In experiments involving T_R depletion, mice were immunized on day 3 after administration of anti-CD25 antibody.

Cell purification. Spleen and draining lymph nodes were harvested from 8- to 12-week-old mice. Single cell suspensions were incubated with anti-CD4 microbeads followed by purification using an AutoMACS sorter (Miltenyi Biotec). In all the experiments more than 95% of recovered cells were CD4 positive. Anti-CD19 microbeads were used to purify B cells from spleen. For experiments involving cell transfer, B cells were purified by incubating single cell suspensions of spleen with a cocktail of anti-CD4, anti-CD8, anti-Thy1 and anti-Mac-1 antibodies followed by rabbit complement (Cedarlane laboratories). Purity of B cells was confirmed by staining with anti-B220 antibody and was always more than 98%.

T-cell proliferation assay. Purified CD4⁺ T cells (1 × 10⁵) from draining lymph nodes were cultured in flat-bottom 96-well plates with irradiated (1,200 rad) B cells (3 × 10⁵) and titrating doses of antigen for 72–84 h. Proliferation of T cells was determined by incorporation of [³H]thymidine for the last 12–16 h of culture.

Enzyme-linked immunosorbent assay. Immunized mice were bled between day 12 and 14 and antigen-specific enzyme-linked immunosorbent assays (ELISAs) were done to determine total and isotype-specific antibodies using specific detection reagents.

Staining and flow cytometry. Cells were stained with the relevant antibodies for 30 min on ice and were washed and analysed using a FACScan flowcytometer (BD Biosciences). Data were analysed using FlowJo software (Treestar Inc.).

Immunohistochemistry. Frozen spleens were cut into 6-µm sections and stained with anti-IgM HRP (Jackson ImmunoResearch) and PNA-biotin (Vector Labs) followed by streptavidin-alkaline phosphates (BD biosciences). Slides were developed with diaminobenzidine and New Fuchsin substrates (BD biosciences).

Quantitative PCR. Total RNA was isolated from cells using Trizol (Invitrogen), and cDNA was synthesized. Real-time PCR analysis was performed using an Mx3000p instrument (Stratagene) to measure SyBR green (Quantitect from Qiagen) incorporation. Relative amounts of mRNA were normalized to HPRT RNA levels within each sample.

Received 4 August; accepted 29 September 2005.

- Janeway, C. A. Jr Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).
- Janeway, C. A. Jr & Medzhitov, R. Innate immune recognition. *Annu. Rev. Immunol.* **20**, 197–216 (2002).
- Takeda, K., Kaisho, T. & Akira, S. Toll-like receptors. *Annu. Rev. Immunol.* **21**, 335–376 (2003).
- Banchereau, J. & Steinman, R. M. Dendritic cells and the control of immunity. *Nature* **392**, 245–252 (1998).
- Kawai, T., Adachi, O., Ogawa, T., Takeda, K. & Akira, S. Unresponsiveness of MyD88-deficient mice to endotoxin. *Immunity* **11**, 115–122 (1999).
- Schnare, M. et al. Toll-like receptors control activation of adaptive immune responses. *Nature Immunol.* **2**, 947–950 (2001).
- Pasare, C. & Medzhitov, R. Toll-dependent control mechanisms of CD4 T cell activation. *Immunity* **21**, 733–741 (2004).
- Kitamura, D., Roes, J., Kuhn, R. & Rajewsky, K. A B cell-deficient mouse by targeted disruption of the membrane exon of the immunoglobulin µ chain gene. *Nature* **350**, 423–426 (1991).

- Kawabe, T. et al. The immune responses in CD40-deficient mice: impaired immunoglobulin class switching and germinal center formation. *Immunity* **1**, 167–178 (1994).
- DiSanto, J. P., Bonnefoy, J. Y., Gauchat, J. F., Fischer, A. & de Saint Basile, G. CD40 ligand mutations in x-linked immunodeficiency with hyper-IgM. *Nature* **361**, 541–543 (1993).
- Hayashi, F. et al. The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
- Bondada, S., Wu, H., Robertson, D. A. & Chelvarajan, R. L. Accessory cell defect in unresponsiveness of neonates and aged to polysaccharide vaccines. *Vaccine* **19**, 557–565 (2000).
- Vos, Q., Lees, A., Wu, Z. Q., Snapper, C. M. & Mond, J. J. B-cell activation by T-cell-independent type 2 antigens as an integral part of the humoral immune response to pathogenic microorganisms. *Immunol. Rev.* **176**, 154–170 (2000).
- McHeyzer-Williams, L. J. & McHeyzer-Williams, M. G. Antigen-specific memory B cell development. *Annu. Rev. Immunol.* **23**, 487–513 (2005).
- Rajewsky, K. Clonal selection and learning in the antibody system. *Nature* **381**, 751–758 (1996).
- Goodnow, C. C. et al. Altered immunoglobulin expression and functional silencing of self-reactive B lymphocytes in transgenic mice. *Nature* **334**, 676–682 (1988).
- Laszlo, G., Hathcock, K. S., Dickler, H. B. & Hodes, R. J. Characterization of a novel cell-surface molecule expressed on subpopulations of activated T and B cells. *J. Immunol.* **150**, 5252–5262 (1993).
- Dent, A. L., Shaffer, A. L., Yu, X., Allman, D. & Staudt, L. M. Control of inflammation, cytokine expression, and germinal center formation by BCL-6. *Science* **276**, 589–592 (1997).
- Shaffer, A. L. et al. BCL-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control. *Immunity* **13**, 199–212 (2000).
- Angelini-Duclos, C., Cattoretti, G., Lin, K. I. & Calame, K. Commitment of B lymphocytes to a plasma cell fate is associated with Blimp-1 expression *in vivo*. *J. Immunol.* **165**, 5462–5471 (2000).
- Turner, C. A. Jr, Mack, D. H. & Davis, M. M. Blimp-1, a novel zinc finger-containing protein that can drive the maturation of B lymphocytes into immunoglobulin-secreting cells. *Cell* **77**, 297–306 (1994).
- Schliephake, D. E. & Schimpl, A. Blimp-1 overcomes the block in IgM secretion in lipopolysaccharide/anti-µ F(ab')₂-co-stimulated B lymphocytes. *Eur. J. Immunol.* **26**, 268–271 (1996).
- Soro, P. G. et al. Differential involvement of the transcription factor Blimp-1 in T cell-independent and -dependent B cell differentiation to plasma cells. *J. Immunol.* **163**, 611–617 (1999).
- Shaffer, A. L. et al. Blimp-1 orchestrates plasma cell differentiation by extinguishing the mature B cell gene expression program. *Immunity* **17**, 51–62 (2002).
- Shapiro-Shelef, M. et al. Blimp-1 is required for the formation of immunoglobulin secreting plasma cells and pre-plasma memory B cells. *Immunity* **19**, 607–620 (2003).
- Bernasconi, N. L., Traggiai, E. & Lanzavecchia, A. Maintenance of serological memory by polyclonal activation of human memory B cells. *Science* **298**, 2199–2202 (2002).
- Barton, G. M. & Medzhitov, R. Control of adaptive immune responses by Toll-like receptors. *Curr. Opin. Immunol.* **14**, 380–383 (2002).
- Leadbetter, E. A. et al. Chromatin-IgG complexes activate B cells by dual engagement of IgM and Toll-like receptors. *Nature* **416**, 603–607 (2002).
- Rui, L., Vinuesa, C. G., Blasioli, J. & Goodnow, C. C. Resistance to CpG DNA-induced autoimmunity through tolerogenic B cell antigen receptor ERK signalling. *Nature Immunol.* **4**, 594–600 (2003).
- Moller, G., Andersson, J. & Sjöberg, O. Lipopolysaccharides can convert heterologous red cells into thymus-independent antigens. *Cell. Immunol.* **4**, 416–424 (1972).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by a grant from the NIH. R.M. is an investigator of the Howard Hughes Medical Institute. We thank L. Kopp for providing flagellin, N. Palm for help with analysing the CD11c-MyD88 transgenic mice, and S. Akira for providing MyD88- and TLR4-deficient mice.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.M. (ruslan.medzhitov@yale.edu).

Evidence for *de novo* imprinted X-chromosome inactivation independent of meiotic inactivation in mice

Ikuhiro Okamoto¹, Danielle Arnaud², Patricia Le Baccon¹, Arie P. Otte³, Christine M. Distèche⁴, Philip Avner² & Edith Heard¹

In mammals, one of the two X chromosomes is inactivated in females to enable dosage compensation for X-linked gene products¹. In rodents and marsupials, only the X chromosome of paternal origin (Xp) is silenced during early embryogenesis. This could be due to a carry-over effect of the X chromosome's passage through the male germ line, where it becomes transiently silenced together with the Y chromosome, during meiotic sex chromosome inactivation (MSCI)². Here we show that *Xist* (X inactive specific transcript) transgenes, located on autosomes, do not undergo MSCI in the male germ line of mice and yet can induce imprinted *cis*-inactivation when paternally inherited, with identical kinetics to the Xp chromosome. This suggests that MSCI is not necessary for imprinted X-chromosome inactivation in mice. We also show that the Xp is transcribed, like autosomes, at zygotic gene activation rather than being 'pre-inactivated'³. We propose that expression of the paternal *Xist* gene at zygotic gene activation is sufficient to trigger *cis*-inactivation of the X chromosome, or of an autosome carrying a *Xist* transgene.

In some mammals, only the X chromosome of paternal origin is inactivated during early female pre-implantation embryogenesis. In mice, imprinted Xp inactivation is maintained in the trophectoderm⁴, but is reversed in the inner cell mass of the blastocyst, before random X-chromosome inactivation (X inactivation) in embryonic tissues^{5,6}. Imprinted X inactivation may be the ancestral form of the process, as it occurs in all tissues of marsupials^{7,8}. The nature of the imprint(s) underlying Xp inactivation is the subject of debate. One theory involved a mark that prevents inactivation of the maternal X chromosome (Xm) during early embryogenesis⁹. The early death of mouse embryos carrying two Xm chromosomes, presumably due to the lack of dosage compensation, supports this^{10,11}. The maternal imprint responsible is acquired during oocyte maturation¹² and may act by repressing *Xist* until the late morula stage¹³ (Fig. 1a). Another theory posits that imprinted Xp inactivation results from a predisposition of the paternal X chromosome for inactivation^{14–15}, and even that the Xp is inherited in a pre-inactivated state that is already dosage compensated after fertilization³, as a result of its passage through the male germ line where it associates with the Y chromosome, forming the XY body, and undergoing MSCI at the pachytene stage of meiosis² (Fig. 1a).

To distinguish between these theories and to define the X-chromosomal region carrying the imprint regulating this process, we have examined imprinted X inactivation in mice carrying large, single-copy, *Xist*-containing yeast artificial chromosome (YAC) transgenes located on autosomes. Previous studies, aimed at defining the

minimal X-inactivation centre (Xic) region sufficient to induce ectopic X inactivation, had shown that unlike multicopy arrays, such single-copy transgenes could not induce random X inactivation in mouse somatic tissues and differentiating embryonic stem (ES) cells^{16,17}. However, their capacity to induce imprinted inactivation during development had not been examined. Furthermore, although both imprinted and random X inactivation are *Xist*-dependent^{18,19}, the extent of the region flanking *Xist* that is necessary and sufficient to induce imprinted X inactivation is unknown. We examined two single-copy transgenic mouse lines—Tg53 and Tg80 carrying 460-kilobase (kb) and 210-kb *Xist*-containing fragments, respectively¹⁶ (Fig. 1b)—for their capacity to induce imprinted *cis*-inactivation. Using RNA fluorescence *in situ* hybridization (FISH), *Xist* expression profiles were examined in blastocysts derived from males or females hemizygous for the transgene, which had been crossed with non-transgenic partners (Fig. 1c, d). Early female (XmXp) blastocysts normally show a single *Xist* RNA accumulation on the Xp chromosome in all blastomeres. Male (XmY) blastocysts never show a *Xist* RNA domain and the Xm chromosome displays a punctate *Xist* RNA signal corresponding to low-level *Xist* expression in some blastomeres. The results obtained using embryos derived from transgenic mice (Supplementary Information) demonstrate that autosomal *Xist* transgenes show normal patterns of imprinted *Xist* expression: maternally inherited alleles are repressed (Fig. 1d), displaying only a punctate *Xist* RNA signal in blastocysts, whereas paternally inherited alleles are expressed and accumulate *Xist* transcript *in cis* over the autosome that carries them. In the latter, transgenic female embryos show the unique pattern of two *Xist* RNA domains in all blastomeres, one corresponding to the Xp and the other to the transgenic autosome. A more detailed description of these data are provided in the Supplementary Information and Supplementary Table 1. On the basis of this, we conclude that the region that is necessary and sufficient to impose the imprint(s) underlying monoallelic *Xist* expression in early embryos must lie within these transgenes (Fig. 1b).

Next, we assessed whether paternally inherited *Xist* transgenes can induce chromatin modifications associated with imprinted Xp inactivation in pre-implantation embryos⁶. Hypomethylation of H3 K4 and hypoacetylation of H3 K9 and H4 (data not shown) were found at both the Xp and transgenic *Xist* RNA domains in female embryos carrying a paternally inherited Tg53 transgene, from the 8-cell through to the blastocyst stage. Enrichment of the Polycomb group proteins Ezh2 and Eed, and associated H3 K27 trimethylation, could be detected from the 16–32-cell stage onwards at the site of both the Xp and the Tg53- or Tg80-associated autosomes

¹CNRS UMR218, Curie Institute, 26 rue d'Ulm, 75248 Paris Cedex 05, France. ²Pasteur Institute, 25 rue du Docteur Roux, Paris 75015, France. ³Swammerdam Institute for Life Sciences, University of Amsterdam, Kruislaan 406, 1098 SM Amsterdam, The Netherlands. ⁴Department of Pathology, University of Washington, Seattle, Washington 98195, USA.

(Fig. 2; see also Supplementary Figs 1 and 2). On the other hand, maternally transmitted Tg53 transgenes did not show *Xist* RNA and H3 tri-methyl K27 accumulation at the morula stage (Supplementary Fig. 1b). We also investigated the replication timing of the paternally inherited Tg53-linked autosome (chromosome 13). Early (rather than late) replication is a classical trait of imprinted Xp inactivation (see Supplementary Methods). An asynchronously (early) replicating chromosome 13 was detected in paternally derived Tg53 blastocysts (Fig. 2d), but never in wild-type blastocysts, indicating that Tg53 had indeed resulted in a shift in replication timing. Thus, paternally inherited *Xist* transgenes are capable of inducing multiple characteristics of imprinted inactivation *in cis* during pre-implantation embryogenesis, with identical kinetics to the Xp chromosome, despite their location on autosomes. The maintenance of this

imprinted pattern of inactivation during post-implantation development will require further studies.

Transcriptional activity of Xp or Tg53-associated chromatin was assessed, at the earliest stages of embryogenesis, using three different assays. In the pre-inactivation hypothesis³, the expression of *Xist* is not a pre-requisite for the initial inactivity of the Xp in the female embryo. Rather, it is proposed that passage through the XY body and MSCI ensures this and that the Xp is already dosage compensated when zygotic gene activation occurs³. We examined embryos at the 2-cell stage, which is when major zygotic transcription occurs, for evidence of pre-inactivation. The onset of paternal *Xist* expression was detected as a small dot in 2-cell embryos, whether for the X-linked or transgenic locus (Fig. 3a, b). The consistently small size of this *Xist* RNA signal clearly rules out that the RNA is coating all

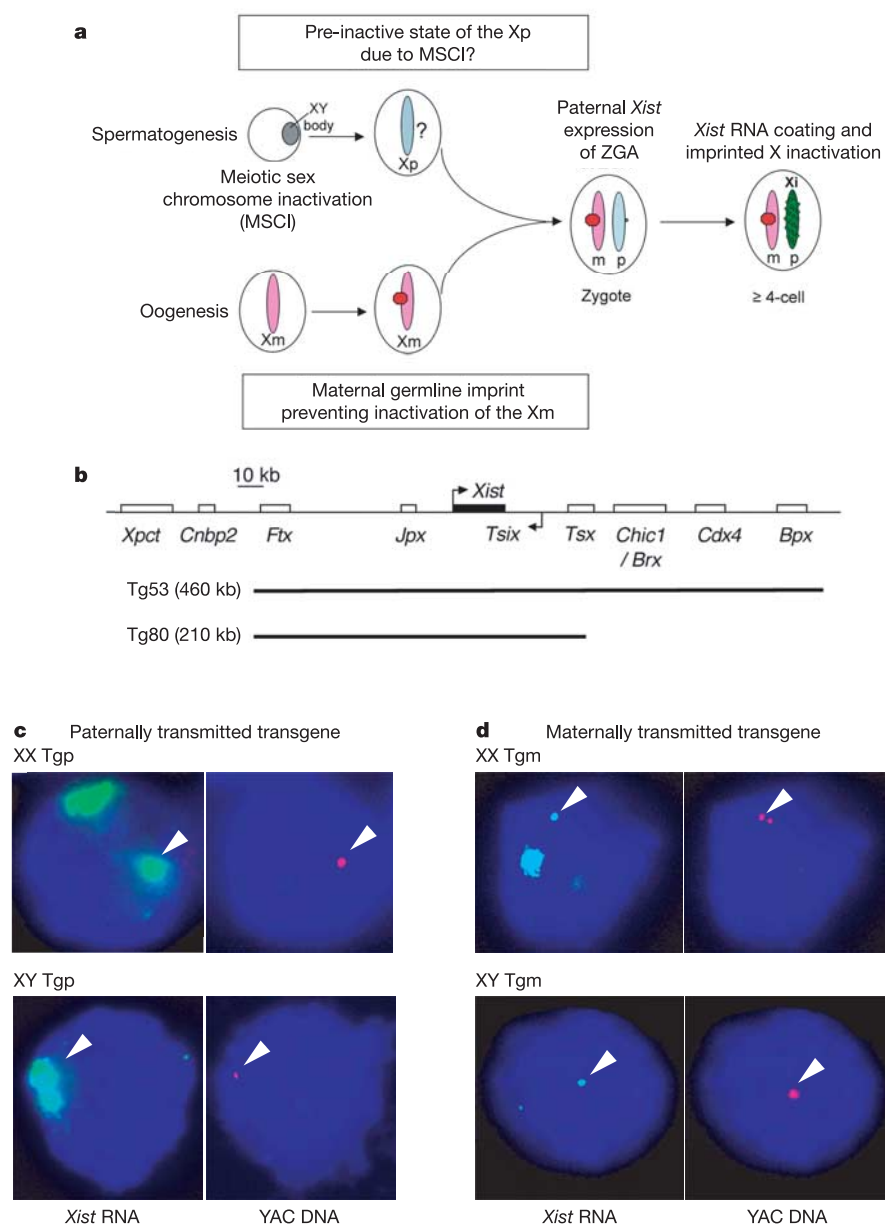


Figure 1 | Imprinted X-inactivation and autosomal *Xist* transgenes.

a, Germline imprints potentially underlying paternal X inactivation. A maternal imprint prevents the maternal X chromosome from undergoing inactivation in early female embryos and/or inactivity of the paternal X chromosome is carried over into the zygote after MSCI in the male germ line. ZGA, zygotic gene activation. **b**, Map of the regions covered by the two

Xist YAC transgenes^{16,17}. **c**, *Xist* RNA FISH (green) was followed by YAC vector DNA FISH (red) to enable transgene detection in blastocysts derived from Tg53 male mice (Tg80 shows similar results). **d**, Same as for **c** except that blastocysts were derived from Tg53 female mice (Tg80 shows similar results).

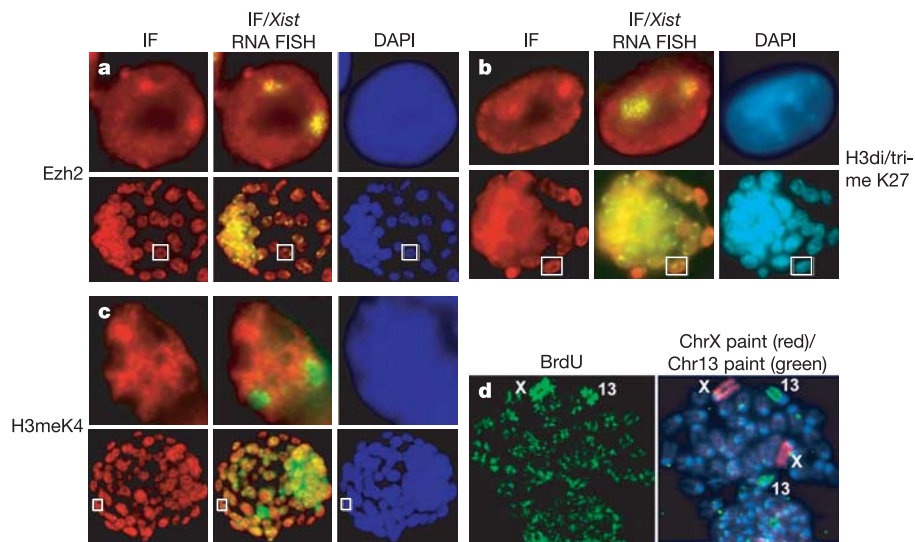


Figure 2 | Histone H3 modifications and Polycomb proteins in female embryos carrying paternally transmitted Tg53 *Xist* transgenes.

a–c, Immunolabelling (red) with antibodies against Ezh2 (**a**), H3 di/tri-methyl K27 (**b**) and H3 di-methyl K4 (**c**) was combined with *Xist* RNA FISH (green). For each stage, an intact embryo together with an enlarged

representative nucleus is shown. **d**, Detection of early replicating chromosomes (X or 13) in blastocysts derived from females crossed with Tg53 males. Immunolabelling (green) with antibodies against BrdU was followed by DNA FISH using X chromosome (red) and chromosome 13 (pseudo-coloured green) paint probes. IF, immunofluorescence.

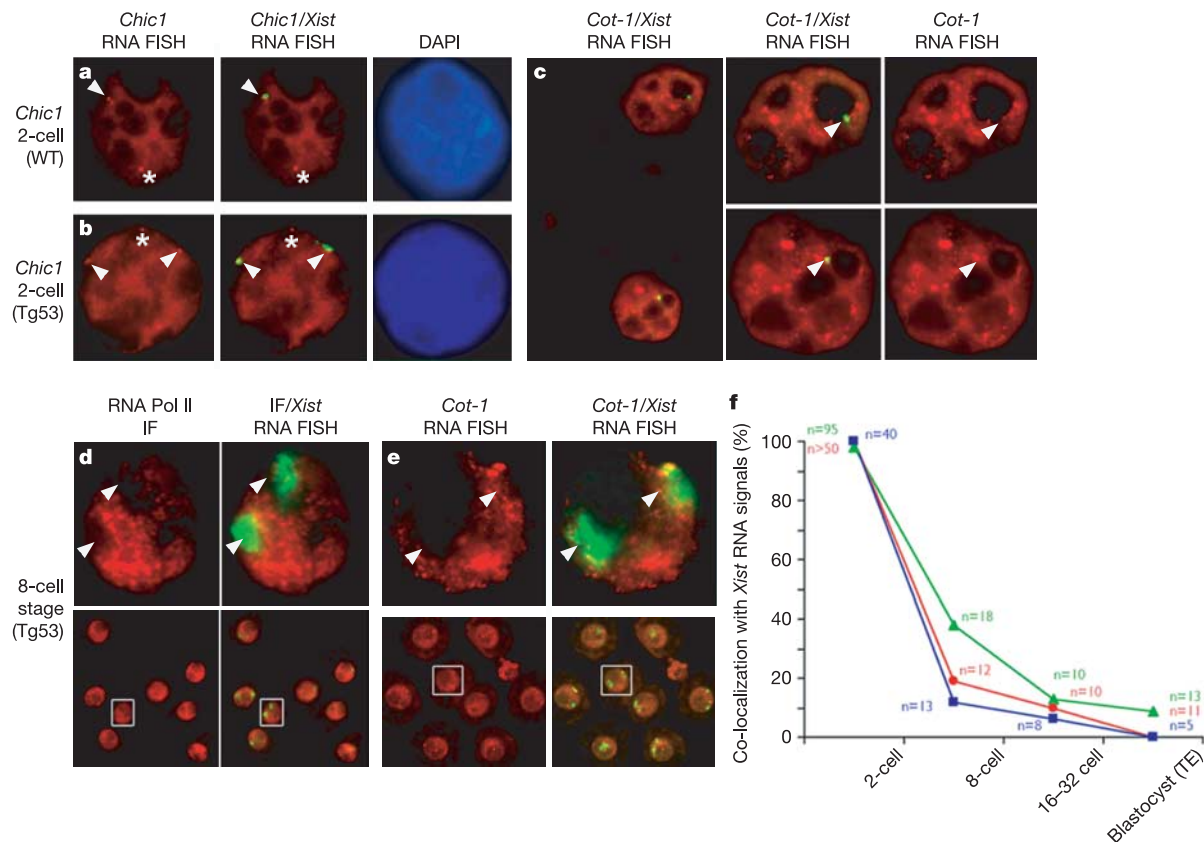


Figure 3 | Early transcriptional activity followed by silencing during imprinted X inactivation. **a, b**, *Chic1* (red) and *Xist* (green) RNA FISH on blastomeres of wild-type (WT) and transgenic (Tg53) female 2-cell embryos. Paternal *Xist* transcription site (arrowhead) is associated with *Chic1* transcribing; maternal X-linked *Chic1* expression is indicated by an asterisk. **c**, *Cot-1* (red) and *Xist* (green) RNA FISH on a wild-type 2-cell embryo. No *Cot-1* exclusion is detectable at or around the site of the *Xist* signal at this stage. **d**, RNA polymerase II (Pol II; red)/*Xist* RNA (green) on a Tg53 male-derived 8-cell embryo. RNA polymerase II is excluded from both

Xist RNA domains (Xp and Tg53 chromosome 13; see Supplementary Fig. 2) from this stage. **e**, *Cot-1/Xist* RNA FISH on a Tg53 male-derived 8-cell embryo. There is some *Cot-1* exclusion from both *Xist* RNA domains by this stage. **f**, Average percentage of blastomeres showing *Chic1* RNA (green triangles), *Cot-1* RNA (red circles) or RNA polymerase II (blue squares) association with the paternal *Xist* RNA signal in female embryos from the 2-cell stage through to blastocyst. Numbers of embryos examined for each stage are indicated.

of the chromosome, or even a large part of it, as was suggested in a previous study³, and is more in keeping with a transcription focus. Indeed, the *Xist* signal at the 2-cell stage is similar in size to that of other primary transcripts of X-linked genes (see Fig. 3a–c; see also Supplementary Fig. 3). Activity of the Xp or Tg53-associated chromosome was first assayed using *Chic1*, a gene normally subject to both random²⁰ and imprinted³ X inactivation. *Chic1* lies 70 kb 3' to *Xist* and is located within the Tg53 transgene (Fig. 1b), thus enabling us to determine the transcriptional activity of both the transgene and the Xp chromosome simultaneously. Using *Xist/Chic1* RNA FISH, both paternal (transgenic or Xp) and maternal (Xm) *Chic1* transcription foci could be detected in 99% of blastomeres ($n = 190$) at the mid-late 2-cell stage. The paternal *Chic1* allele then becomes silenced in the majority of blastomeres from the 8–16-cell stage onwards (Fig. 3f; see also Supplementary Fig. 3a). Inactivation of the *Chic1* gene at the morula stage is in agreement with the study of ref. 3, but the authors did not examine earlier stages for expression of any X-linked genes. In addition to *Chic1*, we also examined two other X-linked genes—*Smcx* (22 cM distal to *Xist*) and *Ube1x* (36 cM proximal to *Xist*)—for their paternal activity at the 2-cell stage (in wild-type embryos). Both genes have previously been found to be subject to X inactivation^{21,22} even though *Smcx* is partially reactivated later in development²³. In mid-late 2-cell embryos, two primary transcript signals could be detected for each of these genes in 99% of blastomeres ($n = 154$ for *Smcx*; $n = 86$ for *Ube1x*), with one of the two signals located in the vicinity of the paternal *Xist* transcript. These genes therefore appear to be expressed at similar levels from both the paternal and maternal X chromosomes at this stage of development (Supplementary Fig. 3b, c). Paternal *Smcx* expression was examined at 8- and 16-cell stages (data not shown), and showed similar kinetics of X inactivation to *Chic1* (Supplementary Fig. 3a). We also used the *Cot-1* RNA FISH assay previously used by ref. 3 to assess the global transcriptional status of the Xp chromosome in early embryos. To analyse the exact distribution of *Cot-1* RNA at, and around, the *Xist* RNA punctate signal in 2-cell embryos, two different methods of three-dimensional microscopy were used, both of which gave similar results (Supplementary Fig. 4a). Neither the *Xist* locus, nor the chromosomal region surrounding it, were found to be depleted for *Cot-1* RNA signal in 95% of late 2-cell blastomeres ($n = 36$) (Fig. 3c; see also Supplementary Fig. 4a). Similar data were obtained for the paternally inherited transgenic *Xist* locus (Supplementary Fig. 4b). By the 8-cell stage, *Cot-1* RNA was clearly excluded from the *Xist* RNA accumulation at both the Xp and the Tg53-carrying chromosome 13 in some blastomeres; it was excluded from the *Xist* RNA accumulation in the majority of blastomeres by the 16-cell stage (Fig. 3d; see also Supplementary Fig. 2). The Xp chromosome and the transgenic autosome showed identical kinetics of *Cot-1* exclusion from the *Xist* RNA domain. A third assay for transcriptional silencing was the previously reported exclusion of RNA polymerase II from the *Xist* RNA-coated chromosome⁶. RNA polymerase II was excluded from transgenic chromosome 13 with identical kinetics to the Xp chromosome (Fig. 3d; see also Supplementary Fig. 2). All of the above assays indicate that the Xp chromosome and an autosome carrying a paternally inherited *Xist* transgene are transcriptionally active after fertilization, but become inactivated after *Xist* RNA accumulation.

The above findings demonstrate that imprinted *cis*-inactivation is efficiently induced by *Xist* transgenes, in a fashion apparently equivalent to imprinted Xp inactivation. We wished to determine whether MSCI might underlie the capacity of these transgenes to induce imprinted paternal inactivation after fertilization. MSCI normally only affects the X and Y chromosomes and is thought to be triggered by their largely unpaired states during meiotic prophase. Recently, however, it was shown that unsynapsed autosomal regions can also be subject to silencing in the germ line²⁴. We therefore analysed pachytene cells in the testes of hemizygous Tg53 male mice for chromatin marks associated with the XY body and MSCI: namely

H2AX phosphorylation²⁵, macroH2A²⁶, ATR²⁷ and ubiquitinated H2A²⁸. Clear XY body enrichment for these marks was observed in pachytene spermatocytes (Fig. 4, Supplementary Fig. 5 and data not shown). However, neither the chromosome 13 pair, nor the transgene itself, showed any sign of MSCI marks or XY body association ($n > 100$) (Fig. 4). Furthermore, the transgene was transcriptionally active, as a *Chic1* transcript signal of transgenic origin (Fig. 4e, f) could be detected in spermatocytes of transgenic males ($n > 100$). No *Chic1* expression was detectable from the XY body in wild-type or transgenic spermatocytes. Finally, in embryos derived from males homozygous for the transgene (which should permit meiotic pairing), imprinted inactivation of the transgenic autosome showed identical kinetics to the Xp and to the transgenic autosome from a

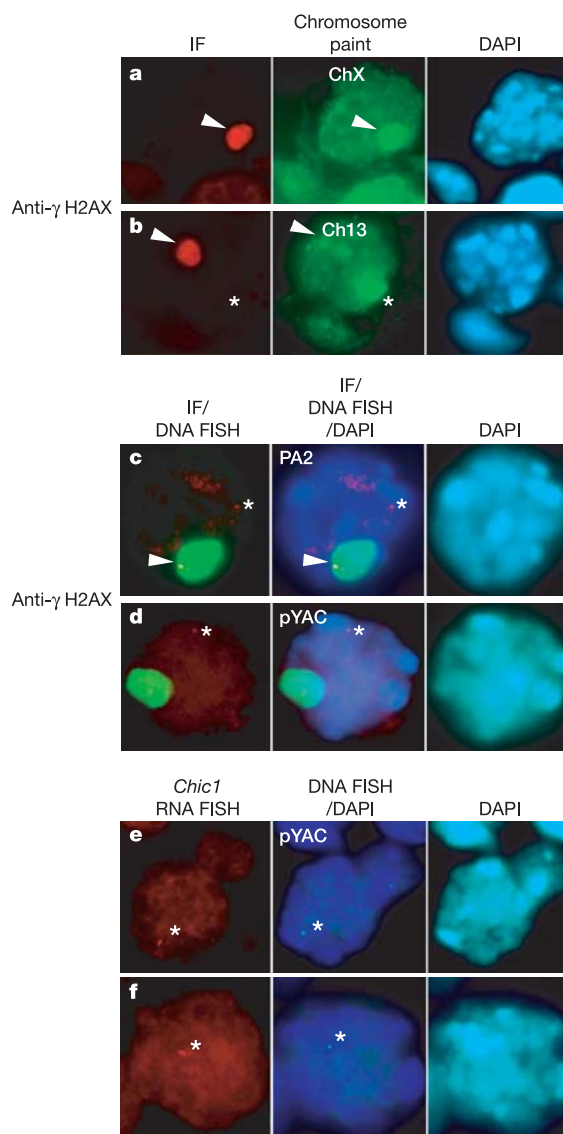


Figure 4 | Absence of XY body marks and meiotic inactivation of the Tg53 transgene on chromosome 13 in pachytene-stage spermatocytes from male Tg53 testes. **a, b**, Immunolabelling (red) with an antibody against the phosphorylated form of H2AX and DNA FISH for the chromosome X or 13 paint (green), on spermatocytes of male Tg53 testes. **c, d**, Immunolabelling (green) with an antibody against phospho-H2AX and DNA FISH with a probe detecting both the endogenous and transgenic Xic loci (YAC PA2) (red) or the transgenic locus only (pYAC). Arrowheads indicate XY body; asterisks indicate chromosome 13 pair and the transgene. **e, f**, *Chic1* RNA FISH (red) on Tg53 spermatocytes (identified based on morphology), followed by DNA FISH with a probe detecting the transgenic locus only (pYAC) (green).

hemizygous father (Supplementary Fig. 6). Thus, the meiotic pairing status of the transgene has no impact on its capacity to induce imprinted *cis*-inactivation after fertilization.

We provide evidence that, in mice, passage through the XY body is not a pre-requisite for imprinted *cis*-inactivation, and that mono-allelic (paternal) *Xist* expression is likely to be the critical determinant for imprinted Xp inactivation²⁹. An autosome carrying a *Xist* YAC transgene behaves similarly to the X chromosome simply because both carry the *Xist* gene surrounded by sufficient regulatory elements to ensure correct imprinted expression patterns for *Xist*. We find no evidence for Xp pre-inactivation³ at the time of zygotic gene activation. Although we cannot exclude that some X chromosomal or transgenic autosomal genes are silent at the 2-cell stage, we can conclude that this is unlikely to be due to meiotic inactivation given our results with autosomal *Xist* transgenes. Our work suggests that, in the face of the silence of the maternal *Xist* gene, the early expression of a paternally inherited *Xist* allele at zygotic gene activation, whether X-linked or autosomal, leads to silencing after *Xist* RNA accumulation at the 4-cell stage. Indeed, the chromatin remodelling of the paternal genome, associated with protamine-histone replacement immediately after fertilization, might even facilitate early transcription of the paternal *Xist* allele³⁰. Whether MSCI of the Xp chromosome played a role in the ancestral form of imprinted X inactivation, which is thought to be represented by marsupials, remains an important question for the future (see Supplementary Discussion). Finally, our study also demonstrates that the Xic sequence requirements for initiating imprinted X inactivation are different compared with those for random inactivation, as the very same single copy *Xist* transgenes studied here cannot induce *Xist* RNA coating and *cis*-inactivation in differentiating ES cells¹⁷ or in somatic tissues of mice¹⁶.

METHODS

Mice and collection of embryos. Transgenic mouse lines Tg53 and Tg80 have previously been described¹⁶. The Tg53 transgene covers a 460-kb region, integrated as a single copy in the pericentric region of chromosome 13. The Tg80 transgene covers a 210-kb region integrated as a single copy in the distal region of an unknown autosome. To obtain hemizygous embryos carrying the transgene either on the paternally or maternally derived autosome we mated F₁ (C57BL/6 × DBA) females with transgenic males or transgenic females with F₁ (C57BL/6 × DBA) males, respectively. Embryo collection and preparation were as described previously⁶. In some cases (in particular for replication assays) blastocysts were cultured for several hours.

Immunofluorescence and RNA/DNA FISH on embryos, testes and ear fibroblasts. Antibodies used are listed in Supplementary Table 1. Details of probes, protocols and microscopy parameters can be found in Supplementary Data.

Received 12 May; accepted 22 August 2005.

Published online 16 October 2005.

1. Lyon, M. F. Gene action in the X chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
2. Lifshytz, E. & Lindsley, D. L. The role of X-chromosome inactivation during spermatogenesis (*Drosophila*-allocyclic-chromosome evolution-male sterility-dosage compensation). *Proc. Natl Acad. Sci. USA* **69**, 182–186 (1972).
3. Huynh, K. D. & Lee, J. T. Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos. *Nature* **426**, 857–862 (2003).
4. Takagi, N. & Sasaki, M. Preferential inactivation of the paternally derived X-chromosome in the extraembryonic membranes of the mouse. *Nature* **256**, 640–642 (1975).
5. Mak, W. *et al.* Reactivation of the paternal X chromosome in early mouse embryos. *Science* **303**, 666–669 (2004).
6. Okamoto, I., Otto, A. P., Allis, C. D., Reinberg, D. & Heard, E. Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science* **303**, 644–649 (2004).
7. Cooper, D. *et al.* X-inactivation in marsupials and monotremes. *Semin. Dev. Biol.* **4**, 117–128 (1993).

8. Vandeberg, J. L., Robinson, E. S., Samollow, P. B. & Johnston, P. G. *Isozymes: Current Topics in Biological and Medical Research* 225–253 (Liss, New York, 1987).
9. Lyon, M. F. & Rastan, S. Parental source of chromosome imprinting and its relevance to X-chromosome inactivation. *Differentiation* **26**, 63–67 (1984).
10. Shao, C. & Takagi, N. An extra maternally derived X chromosome is deleterious to early mouse development. *Development* **110**, 969–975 (1990).
11. Tada, T. & Takagi, N. Parental imprinting on the mouse X chromosome: effects on the early development of XO, XXY and XXX embryos. *Genet. Res.* **62**, 139–148 (1993).
12. Tada, T. *et al.* Imprint switching for non-random X-chromosome inactivation during mouse oocyte growth. *Development* **127**, 3101–3105 (2000).
13. Heard, E. Recent advances in X-chromosome inactivation. *Curr. Opin. Cell Biol.* **16**, 247–255 (2004).
14. Monk, M. & McLaren, A. X-chromosome activity in foetal germ cells of the mouse. *J. Embryol. Exp. Morphol.* **63**, 75–84 (1981).
15. Grant, S. G. & Chapman, V. M. Mechanisms of X-chromosome regulation. *Annu. Rev. Genet.* **22**, 199–233 (1988).
16. Heard, E. *et al.* Transgenic mice carrying an *Xist*-containing YAC. *Hum. Mol. Genet.* **5**, 441–450 (1996).
17. Heard, E., Mongelard, F., Arnaud, D. & Avner, P. *Xist* yeast artificial chromosome transgenes function as X-inactivation centers only in multicopy arrays and not as single copies. *Mol. Cell. Biol.* **19**, 3156–3166 (1999).
18. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for *Xist* in X chromosome inactivation. *Nature* **379**, 131–137 (1996).
19. Marahrens, Y., Panning, B., Dausman, J., Strauss, W. & Jaenisch, R. *Xist*-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev.* **11**, 156–166 (1997).
20. Simmler, M. C. *et al.* Localization and expression analysis of a novel conserved brain expressed transcript, *Brx/BRX*, lying within the Xic/XIC candidate region. *Mamm. Genome* **8**, 760–766 (1997).
21. Carrel, L., Hunt, P. A. & Willard, H. F. Tissue and lineage-specific variation in inactive X chromosome expression of the murine *Smcx* gene. *Hum. Mol. Genet.* **5**, 1361–1366 (1996).
22. Carrel, L. *et al.* X inactivation analysis and DNA methylation studies of the ubiquitin activating enzyme E1 and PCTAIRE-1 genes in human and mouse. *Hum. Mol. Genet.* **5**, 391–401 (1996).
23. Lingenfelter, P. A. *et al.* Escape from X inactivation of *Smcx* is preceded by silencing during mouse development. *Nature Genet.* **18**, 212–223 (1998).
24. Turner, J. *et al.* Silencing of unsynapsed meiotic chromosomes in the mouse. *Nature Genet.* **37**, 41–47 (2005).
25. Fernandez-Capetillo, O. *et al.* H2AX is required for chromatin remodeling and inactivation of sex chromosomes in male mouse meiosis. *Dev. Cell* **4**, 497–513 (2003).
26. Hoyer-Fender, S., Costanzi, C. & Pehrson, J. R. Histone macroH2A1.2 is concentrated in the XY-body by the early pachytene stage of spermatogenesis. *Exp. Cell Res.* **258**, 254–260 (2000).
27. Baart, E. B., de Rooij, D. G., Keegan, K. S. & de Boer, P. Distribution of Atr protein in primary spermatocytes of a mouse chromosomal mutant: a comparison of preparation techniques. *Chromosoma* **109**, 139–147 (2000).
28. Baarends, W. M. *et al.* Histone ubiquitination and chromatin remodeling in mouse spermatogenesis. *Dev. Biol.* **235**, 322–333 (1999).
29. Kay, G. F., Barton, S. C., Surani, M. A. & Rastan, S. Imprinting and X chromosome counting mechanisms determine *Xist* expression in early mouse development. *Cell* **77**, 639–650 (1994).
30. Heard, E., Chaumeil, J., Masui, O. & Okamoto, I. Mammalian X-chromosome inactivation: An epigenetics paradigm. *Cold Spring Harb. Symp. Quant. Biol.* **69**, 89–102 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank G. Almouzni, P. A. Defossez and N. Takagi for critical reading of the manuscript; members of our group and V. Colot for discussions; D. Reinberg and D. Papadopoulos for gifts of antibodies; and M. Blanche for help with mice. I.O. was supported by the CNRS; E.H. and I.O. were financed by the CNRS (ATIP programme), the French Ministry of Research (Action Concertée Incitative), the Curie Institute (Program Incitatif et Collaboratif) and the EU Network of Excellence (Epigenome). P.A. was supported by contracts from ARC by the Ministry of Research under the ACI programme and the EU Network of Excellence (Epigenome).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.H. (Edith.Heard@curie.fr).

LETTERS

A histone H3 methyltransferase controls epigenetic events required for meiotic prophase

Katsuhiko Hayashi^{1,2,†}, Kayo Yoshida³ & Yasuhisa Matsui^{1,2,4}

Epigenetic modifications of histones regulate gene expression and chromatin structure^{1,2}. Here we show that *Meisetz* (meiosis-induced factor containing a PR/SET domain and zinc-finger motif) is a histone methyltransferase that is important for the progression of early meiotic prophase. *Meisetz* transcripts are detected only in germ cells entering meiotic prophase in female fetal gonads and in postnatal testis. Notably, *Meisetz* has catalytic activity for trimethylation, but not mono- or dimethylation, of lysine 4 of histone H3, and a transactivation activity that depends on its methylation activity. Mice in which the *Meisetz* gene is disrupted show sterility in both sexes due to severe impairment of the double-stranded break repair pathway, deficient pairing of homologous chromosomes and impaired sex body formation. In *Meisetz*-deficient testis, trimethylation of lysine 4 of histone H3 is attenuated and meiotic gene transcription is altered. These findings indicate that meiosis-specific epigenetic events in mammals are crucial for proper meiotic progression.

In sexual reproduction, meiosis reduces the ploidy of the genome and generates genomic diversity by shuffling information between homologous chromosomes. To accomplish meiosis, the transcription

of several meiotic genes must be properly orchestrated over time as meiosis progresses. Transcriptional control of gene expression depends crucially on DNA accessibility, which is epigenetically regulated by histone modification^{1,2}. The methylation of lysine 4 of histone H3 (H3K4 methylation) is a well-characterized feature of transcriptionally active genes^{3–6}, indicating that the action of histone methyltransferase (HMTase) on H3K4 marks genes for transcriptional activation according to specific tissue and temporal patterns. Although HMTases that catalyse H3K4 methylation have been identified in mammals^{7–12}, it remains unclear how the epigenetic modification is regulated during meiosis.

To elucidate transcriptional factors controlling the initiation and progression of meiosis, we identified genes whose expression was increased during entry into meiosis by subtracting complementary DNAs of mitotic primordial germ cells at embryonic day 11.5 (E11.5) from those of meiotic female primordial germ cells at E13.5. Of the genes identified (Supplementary Fig. S1), one encoded a putative transcription factor that we named *Meisetz*. The deduced amino acid sequence of *Meisetz* has a PR/SET domain (the catalytic domain of HMTases) in its amino-terminal portion and a C2H2-type zinc-finger

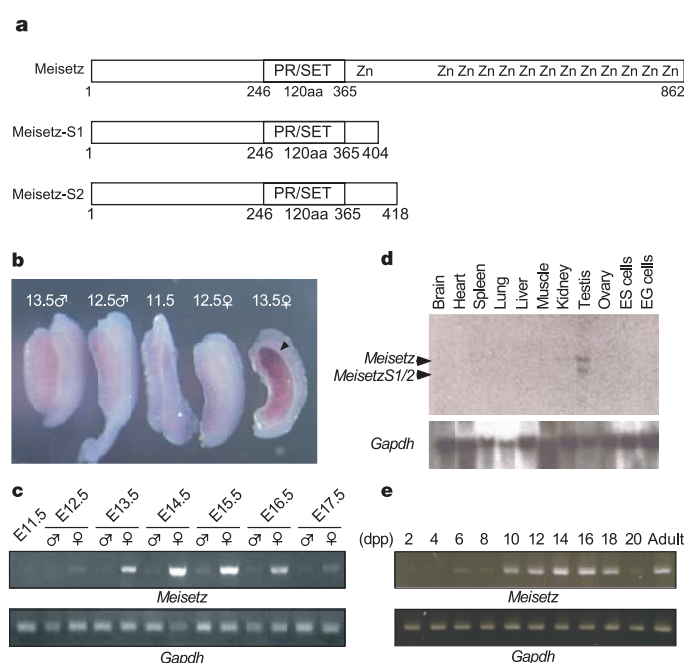


Figure 1 | Molecular structure and expression of the *Meisetz* gene. **a**, Domains in the deduced sequences of *Meisetz* protein and its short isoforms. Numbers indicate the amino acid sequence positions of each domain. **b**, *In situ* hybridization analysis of *Meisetz* expression in fetal gonads. *Meisetz* transcripts were detected in only E13.5 female gonad (arrowhead). **c**, RT-PCR analysis of *Meisetz* expression in fetal gonads. cDNAs obtained from fetal gonads at the indicated developmental stages were amplified with primers for *Meisetz* and *Gapdh*. **d**, Northern blot analysis of *Meisetz* expression in adult tissues. *Gapdh* is shown as a control. **e**, RT-PCR analysis of *Meisetz* expression in the first round of spermatogenesis. cDNAs obtained from testis at the indicated days after birth were amplified with primers for *Meisetz* and *Gapdh*.

¹Department of Molecular Embryology, Research Institute, Osaka Medical Center for Maternal and Child Health, Murodo-cho 840, Izumi, Osaka 594-1101, Japan. ²CREST, Japan Science and Technology Agency (JST), Saitama 332-0012, Japan. ³Department of Molecular Genetics, Graduate School of Medicine, Osaka City University, 1-4-3 Asahimachi, Abeno-ku, Osaka 545-8585, Japan. ⁴Cell Resource Center for Biomedical Research, Institute of Development, Aging and Cancer, Tohoku University, Seiryomachi 4-1, Sendai 980-8575, Japan. [†]Present address: Wellcome Trust/Cancer Research UK, Gurdon Institute of Cancer and Developmental Biology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK.

motif in its carboxy-terminal portion (Fig. 1a and Supplementary Fig. S2a). The amino acid sequence of the PR/SET domain of Meisetz is highly conserved in human (GenBank PRDM7, GenBank PRDM9) and rat (Ensembl ENSRNOG00000021493; Supplementary Fig. S2b). By rapid amplification of cloned ends (RACE), we found two shorter isoforms of the *Meisetz* gene lacking the zinc-finger motif (Fig. 1a and Supplementary Fig. S2a), which are generated by alternative splicing.

Transcripts of *Meisetz* were transiently increased in female gonads from E13.5 to E16.5 (Fig. 1b, c), the time during which meiosis proceeded from pre-meiotic replication to pachytene stages. Its expression was barely detectable in fetal male gonads. In adults, transcripts of *Meisetz* and its splicing variants were detected in testis, but not in any other tissue tested (Fig. 1d). Transcript abundance increased from 10 d post partum (dpp) to 18 dpp (Fig. 1e), during which time the first wave of spermatogenesis proceeds synchronously from pre-leptotene to pachytene stages¹³. The expression patterns of the two isoforms were identical to those of the full-length *Meisetz* gene in fetal gonads and adult testes (data not shown).

We tested whether Meisetz has HMTase activity by an *in vitro* methylation analysis. Recombinant glutathione S-transferase (GST)-fused Meisetz (GST–Meisetz) possessed HMTase activity that specifically catalysed methylation of native histone H3 (native H3) from calf thymus (Fig. 2a). To determine which residue is methylated by Meisetz, we tested the *in vitro* methylation of constructs in which GST was fused to the N terminus of histone H3 protein (GST–H3N) that are suitable for identifying the methylated lysine¹⁴. However, Meisetz did not methylate either GST–H3N or GST-fused full-length H3 (data not shown). Furthermore, Meisetz did not methylate these substrates after the GST tag was removed (data not shown). These results indicate that Meisetz might not catalyse the monomethylation of unmethylated recombinant histone H3, but might preferentially catalyse di- or/and trimethylation of native H3.

To evaluate this possibility, native H3 was incubated with GST–Meisetz and the methylated residues and their methylation status were determined by using methylation-specific antibodies. Unexpectedly, Meisetz showed catalytic activity for only H3K4 trimethylation (Fig. 2b, left). This trimethylation-specific activity of Meisetz was confirmed in experiments showing that Meisetz can restrictively methylate dimethylated H3K4 peptide (Supplementary Fig. S3). Furthermore, overexpression of *Meisetz* in COS7 cells also caused only H3K4 trimethylation (Fig. 2b, right), showing that the catalytic activity is exerted in physiological conditions.

Comparison of Meisetz with other molecules containing a PR/SET domain showed that this domain in Meisetz is distinct from those in established HMTases such as SET7 and SET9 (refs 7, 8), ALR1 (ref. 11), G9a (ref. 14), Suv39h1 and Suv39h2 (refs 15, 16) and SET1 (refs 17, 18; and data not shown). Notably, the PR/SET domain of Meisetz lacks the well-conserved sequence H/RxxNHxC, mutations of which abolish catalytic activity in known HMTases¹⁵; thus, histone methylation by Meisetz is accomplished by an unknown motif in the PR/SET domain.

To identify residues that are crucial for the HMTase activity, deleted or point-mutated variants of Meisetz and its isoform Meisetz-S1 were tested for methylation activity *in vitro*. On the basis of sequence alignment between Meisetz and other PR/SET-containing molecules, we identified Tyr 276 (Y276) and Gly 278 (G278) as potential catalytic residues. Deletion of ten amino acids including these residues abolished the enzymatic activity of Meisetz-S1 (Fig. 2c). Furthermore, replacement of G278, but not Y276, abolished the activity of both Meisetz-S1 and Meisetz (Fig. 2c), showing that G278 is central to the HMTase activity of Meisetz. Consistent with the fact that H3K4 trimethylation is a mark of transcriptionally active genes⁴, wild-type Meisetz promoted transactivation but the replacement of G278 abolished this activity (Fig. 2d). Taken together, these results show that Meisetz possesses HMTase activity specific for H3K4 trimethylation and can activate transcription in a manner dependent on its HMTase activity. To our knowledge, Meisetz is the

first tissue-specific HMTase identified in mammals; this tissue specificity, together with its specificity for H3K4 trimethylation, raises the possibility that Meisetz and structurally related HMTases (Supplementary Fig. S2b) regulate the expression of genes in restricted tissues and/or developmental stages.

To identify functions of Meisetz during meiosis, *Meisetz*^{−/−} mice were generated by targeted disruption of the exon containing the initiation codon, which is shared in all isoforms (Supplementary Fig. S4a–c). No transcripts of either *Meisetz* or its splicing variants were detected in *Meisetz*^{−/−} mice (Supplementary Fig. S4d). Mating of heterozygous (*Meisetz*^{+/-}) mice gave rise to offspring in the expected mendelian distribution (data not shown). *Meisetz*^{−/−} mice were viable and showed no significant alteration in body weight

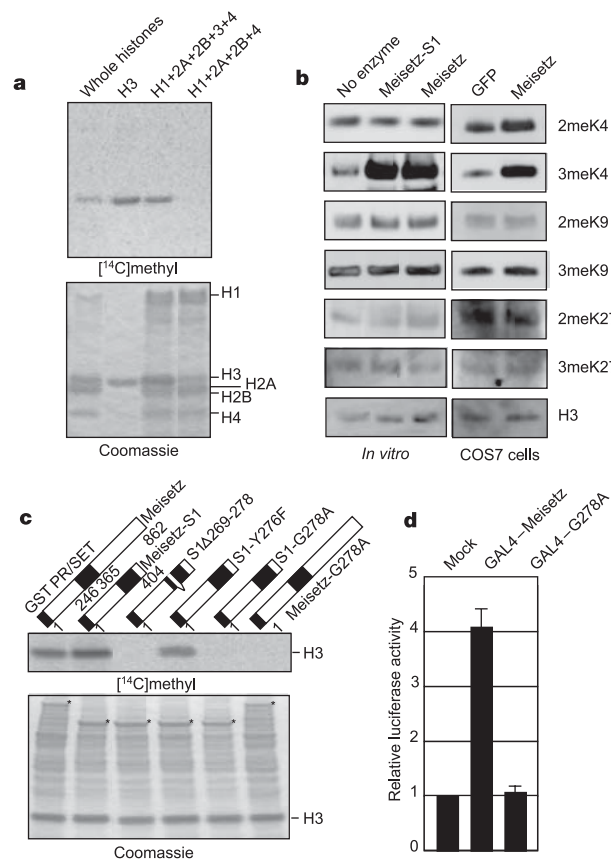


Figure 2 | Meisetz protein catalyses H3K4 trimethylation and activates transcription through its methylation activity. **a**, *In vitro* methylation of native histones by Meisetz. The indicated combinations of native histones were incubated with [¹⁴C]S-adenosyl-L-methionine and GST–Meisetz, separated by SDS–PAGE, stained with Coomassie (bottom), and visualized by autoradiography (top). **b**, Western blot analysis to determine the lysine residue methylated by Meisetz. Native H3 incubated with S-adenosyl-L-methionine and GST–Meisetz or GST–Meisetz-S1 (left), and whole-cell lysates from COS7 cells in which Meisetz or GFP were transiently overexpressed (right), were subjected to western blotting with specific antibodies. **c**, Identification of a key residue for HMTase activity of Meisetz. Native H3 was incubated with [¹⁴C]S-adenosyl-L-methionine and the indicated deletion or point-mutation variants of GST–Meisetz, separated, stained with Coomassie (bottom), and visualized by autoradiography (top). Asterisks indicate the recombinant GST–Meisetz proteins used as enzyme. **d**, Transactivation activity depends on HMTase activity. Expression vector (GAL4–Meisetz or GAL4–G278A) or mock vector was transfected into COS7 cells together with a reporter vector and a control vector. Luciferase activities from the reporter vector were normalized to those from the control vector. Shown are luciferase activities relative to those obtained by transfection of mock vector. Error bars represent the s.d. from three independent experiments.

or appearance (data not shown); however, both sexes of *Meisetz*^{-/-} mice were sterile.

The weight of testes in 8-week-old *Meisetz*^{-/-} mice was nearly 75% less than that of wild type (wild type, 115.3 ± 7.8 mg; *Meisetz*^{-/-}, 29.0 ± 2.2 mg). In 8-week-old *Meisetz*^{-/-} testes, there were no round spermatids, elongated spermatids or spermatozoa (Fig. 3a, b). Using immunofluorescence staining with TRA369 monoclonal antibody, which specifically reacts with both pachytene spermatocytes and spermatids¹⁹, and with a polyclonal antibody against SCP3, a lateral component of synaptonemal complex, we confirmed the accumulation of TRA369⁺ SCP3⁺ pachytene spermatocytes but not TRA369⁺ SCP3⁻ spermatids in *Meisetz*^{-/-} testis (Fig. 3c, d). In *Meisetz*^{-/-} females, there were no growing follicles at any developmental stage in 5-week-old ovaries (Fig. 3e, f), and few germ cells were recognized in neonatal ovaries (Fig. 3i, j). The number of germ cells in gonads decreased from E17.5 onwards (Fig. 3g, h) when meiosis was proceeding through the pachytene stage. These findings show that gametogenesis was disrupted at the pachytene stage in both sexes of *Meisetz*^{-/-} mice.

We examined details of meiotic deficiencies in *Meisetz*^{-/-} testis in spread cells. At the leptotene stage, the distribution of histone H2AX phosphorylated on serine (termed γ H2AX), which marks sites of double-stranded breaks (DSBs)²⁰, showed no significant difference between wild-type and *Meisetz*^{-/-} spermatocytes (Fig. 3k, l). At zygotene and pachytene stages, however, γ H2AX was inappropriately retained on each chromosome in *Meisetz*^{-/-} spermatocytes (Fig. 3n, p), whereas it disappeared from autosomal regions of synapsed chromosomes and converged around the sex chromosomes to form the sex body in wild type²⁰ (Fig. 3m, o). Similar to spermatocytes, female fetal germ cells at the pachytene stage showed more γ H2AX in *Meisetz*^{-/-} than in wild-type mice (Fig. 3q, r). In addition to abnormal retention of γ H2AX, the pachytene chromosomes of *Meisetz*^{-/-} mice were frequently branched and connected

with other chromosomes (Fig. 3p, r), indicating that pairing between homologous chromosomes was also impaired in *Meisetz*^{-/-} mice.

DMC1 protein, which is involved in recombination-associated DNA repair in meiosis^{21,22}, was poorly localized in γ H2AX-stained regions in *Meisetz*^{-/-} spermatocytes (Fig. 3t), whereas in wild-type spermatocytes most of the DMC1 foci were colocalized with γ H2AX signals (Fig. 3s). In addition, the number of DMC1 foci in *Meisetz*^{-/-} spermatocytes was less than half of that in wild type (Fig. 3t, s, and data not shown). Because the level of *Dmc1* transcripts in *Meisetz*^{-/-} testis was comparable to that in wild type (data not shown), DMC1 foci may not be stable unless they properly localize at DSB sites^{23,24}. These results suggest that DSB repair cannot proceed properly in *Meisetz*^{-/-} spermatocytes due to failed localization of DMC1 at DSB sites. Thus, we conclude that meiotic arrest in *Meisetz*^{-/-} spermatocytes is caused by impairment of DSB repair and consequent impairment of pairing between homologous chromosome and sex body formation.

To address further the functional requirement of *Meisetz* in meiosis, we examined the status of H3K4 trimethylation by immunofluorescence in juvenile testes at 14 dpp, when the first wave of spermatogenesis synchronously proceeds to early pachytene. There were no histological differences between wild-type and *Meisetz*^{-/-} testes at this age (data not shown). Signals for H3K4 trimethylation were increased in pachytene spermatocytes in wild-type testis, but this increase was reduced in *Meisetz*^{-/-} testis (Fig. 4a, b). The amount of H3K4 trimethylation in 14-dpp *Dmc1*^{-/-} testes, in which spermatogenesis is arrested at the pachytene stage owing to impairment of chromosomal pairing^{21,22}, was similar to that in wild type (Fig. 4a, b), excluding the possibility that the attenuated H3K4 trimethylation in *Meisetz*^{-/-} testis is a consequence of impaired chromosomal pairing followed by transcriptional repression^{25,26}. The weak increase in H3K4 trimethylation in *Meisetz*^{-/-} pachytene spermatocytes as compared with somatic cells indicates that an

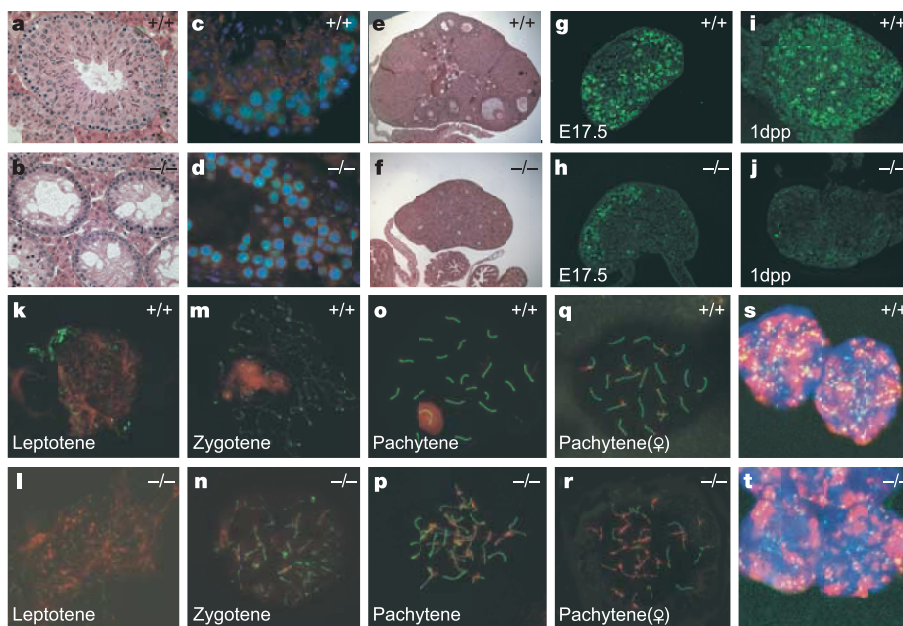


Figure 3 | Impairment of meiotic progression in *Meisetz*^{-/-} mice. Haematoxylin and eosin (HE)-stained sections of wild-type (a) and *Meisetz*^{-/-} (b) testis showed arrest of spermatogenesis in the *Meisetz*^{-/-} testis. Sections of wild-type (c) and *Meisetz*^{-/-} (d) testis stained by anti-SCP3 antibody (green), TRA369 monoclonal antibody (red) and DAPI (blue) confirmed arrest of spermatogenesis at pachytene. HE-stained sections of wild-type (e) and *Meisetz*^{-/-} (f) ovaries, and anti-mVASA (green) immunofluorescence of sections of wild-type (g, i) and *Meisetz*^{-/-}

(h, j) ovaries showed arrest of oogenesis in *Meisetz*^{-/-} ovary. The developmental stages of the ovary are indicated in each panel. Testicular (k–p) and ovarian (q, r) cells were spread and then stained with anti-SCP3 (green) and anti- γ H2AX antibody (red). The developmental stages of meiotic prophase based on SCP3 kinetics are indicated in each panel. Wild-type (s) and *Meisetz*^{-/-} (t) testicular cells were spread and then stained with anti-DMC1 (green) and anti- γ H2AX (red) antibodies and DAPI (blue).

unidentified HMTase also functions in trimethylation. In contrast to H3K4 trimethylation, the amount of H3K4 dimethylation was higher in *Meisetz*^{-/-} testis than in wild type, indicating that dimethylated H3K4 accumulates in *Meisetz*^{-/-} spermatocytes owing to disruption of the H3K4 trimethylation reaction. Taken together, these results indicate that *Meisetz* has essential functions in spermatocytes through its H3K4 trimethylation of a portion of a genome; this H3K4 trimethylation may activate transcription of a group of genes that are crucial for the progression of meiotic prophase.

To address the possibility that an alteration in gene expression is responsible for the meiotic defects in *Meisetz*^{-/-} mice, we screened *Meisetz*^{-/-} mice for genes showing reduced expression by microarray and polymerase chain reaction with reverse transcription (RT-PCR) using cDNA from 14-dpp testes (data not shown). Of the genes identified, one candidate, *4932411A10Rik* (NCBI accession NM177719), was investigated further because its predicted amino acid sequence has a GHKL ATPase motif in its N-terminal region, a motif that is well conserved among DNA mismatch repair, topoisomerase, histidine kinase and HSP90 proteins²⁷ (Supplementary Fig. S5). *4932411A10Rik* transcripts were barely detected in *Meisetz*^{-/-} testis (Fig. 4c), but were expressed in *Dmc1*^{-/-} testis, indicating that the attenuated transcription was caused by loss of *Meisetz* but not by meiotic arrest. H3K4 trimethylation on the transcription start site of

the gene was attenuated (Fig. 4d). In addition, *4932411A10Rik* transcripts were detected only in testis among the adult tissues tested (Fig. 4e) and were transiently expressed from pre-meiotic replication to the early pachytene stage (Fig. 4f, g), corresponding with the time of *Meisetz* expression. Taken together, the results strongly suggest that the *4932411A10Rik* gene is directly regulated by *Meisetz* and is involved in early meiotic progression.

Here we have identified a meiosis-induced histone trimethyltransferase, *Meisetz*, and have shown that it is essential for progression through early meiotic prophase, including DSB repair, homologous chromosome pairing and sex body formation. We propose that *Meisetz* might regulate gene expression through meiosis. Another study has provided evidence that an unknown HMTase for H3K9 also has a role in spermatocytes²⁸. Our results also suggest that an additional HMTase trimethylates H3K4 in pachytene spermatocytes. These facts lead to the possibility that there is a previously unknown group of meiosis-specific HMTases responsible for the proper control of both meiotic gene expression and chromosomal structure, and we propose that *Meisetz* is the first example of such a group of molecules. We plan further studies to clarify how the genome-wide alternation of histone modifications, including trimethylation by *Meisetz*, contributes to the structural control of chromosomal alignment in mammalian meiosis.

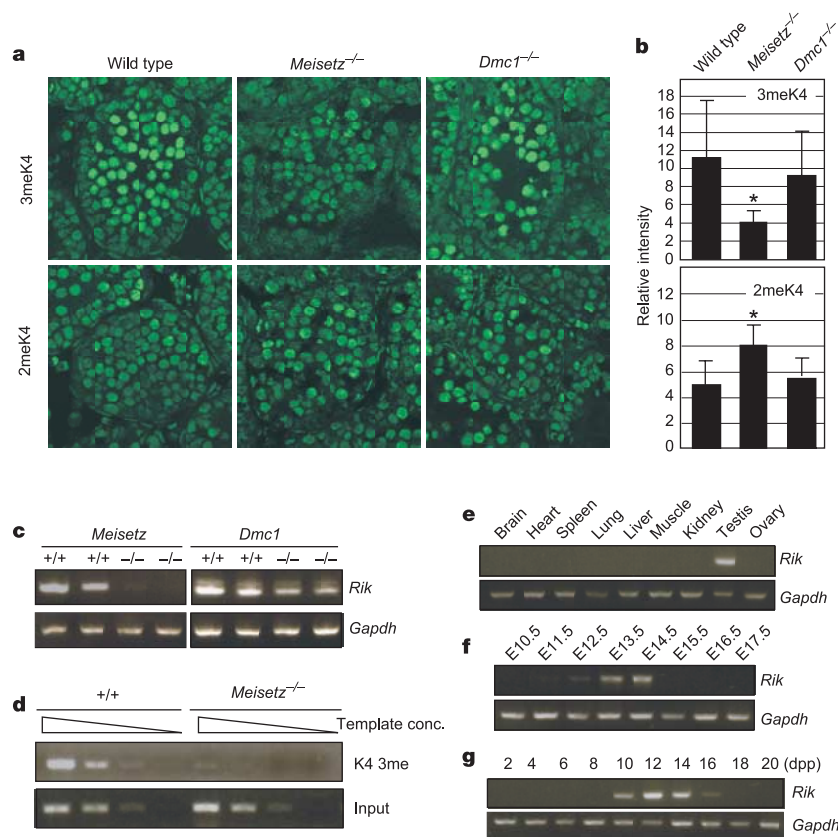


Figure 4 | Downregulation of H3K4 trimethylation and meiotic gene expression in *Meisetz*^{-/-} testis. **a**, Immunofluorescence analysis of H3K4 trimethylation and dimethylation in 14-dpp testis. Sectioned testes of the indicated genotypes were stained with antibodies against trimethylated (3meK4) or dimethylated (2meK4) H3K4. **b**, Relative signal intensity of H3K4 trimethylation and dimethylation. Quantification of the relative intensity of fluorescence signals is described in Supplementary Methods. Asterisks indicate that the difference in signal intensity between *Meisetz*^{-/-} and wild type was statistically significant ($P < 0.05$; Welch's t -test). Error bars indicate the s.d. **c**, Expression of the *4932411A10Rik* gene in *Meisetz*^{-/-} testis. Transcripts of *4932411A10Rik* (*Rik*) or *Gapdh* were amplified by using

cDNAs from 14-dpp testes of the indicated genotypes. **d**, Chromatin immunoprecipitation analysis of a putative transcriptional start site of the *4932411A10Rik* gene. DNA precipitated with antibodies against trimethylated H3K4 (K4 3me) or input DNA derived from testicular cells of each genotype was serially diluted and used for PCR. Additional PCR analyses using three *Meisetz*^{-/-} and two wild-type (+/+) littermates gave similar results (not shown). **e–g**, Meiosis-specific expression of the *4932411A10Rik* gene. Transcripts of *4932411A10Rik* were amplified using cDNA from adult tissues (**e**), fetal female gonads (**f**) and juvenile testes (**g**). Tissues and developmental stages tested are indicated in each panel.

METHODS

Detailed methods are given in the Supplementary Information.

Vectors and methylation and transactivation assays. Full-length and short isoforms of *Meisetz* cDNA were amplified from testis cDNA of C57BL/6N mice by PCR. We amplified the genomic fragments for a targeting vector from the genomic DNA of E14 ES cells. *In vitro* methylation assays were done by using native histones or H3K4 peptides as substrates and GST-fused *Meisetz* as enzymes. For the GAL4 transactivation assay, GAL4-*Meisetz*, GAL4-G278A or mock vector, pFR-luc and pRL-tk were transfected into COS7 cells, and after cultivation cells were lysed and the luciferase activity of each was determined.

Tissue analysis, RT-PCR and chromatin immunoprecipitation. For immunofluorescence analyses of tissues, testes or ovaries were fixed in 4% paraformaldehyde overnight, embedded and sectioned. For RT-PCR analysis, total RNAs from whole gonads of various developmental stages were used for cDNA synthesis by using oligo d(T) primer. Chromatin immunoprecipitation assay was done with dispersed testicular cells. Cells were fixed and sonicated, and the lysate was reacted with rabbit antibodies against trimethylated H3K4. DNA was extracted from the precipitated nucleosomes and subjected to PCR analysis.

Received 20 April; accepted 14 July 2005.

- Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
- Zhang, Y. & Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.* **15**, 2343–2360 (2001).
- Noma, K., Allis, C. D. & Grewal, S. I. Transitions in distinct histone H3 methylation patterns at the heterochromatin domain boundaries. *Science* **293**, 1150–1155 (2001).
- Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407–411 (2002).
- Bernstein, B. E. *et al.* Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl Acad. Sci. USA* **99**, 8695–8700 (2002).
- Schneider, R. *et al.* Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nature Cell Biol.* **6**, 73–77 (2004).
- Wang, H. *et al.* Purification and functional characterization of a histone H3-lysine 4-specific methyltransferase. *Mol. Cell* **8**, 1207–1217 (2001).
- Nishioka, K. *et al.* Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. *Genes Dev.* **16**, 479–489 (2002).
- Milne, T. A. *et al.* MLL targets SET domain methyltransferase activity to Hox gene promoters. *Mol. Cell* **10**, 1107–1117 (2002).
- Nakamura, T. *et al.* ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation. *Mol. Cell* **10**, 1119–1128 (2002).
- Goo, Y. H. *et al.* Activating signal cointegrator 2 belongs to a novel steady-state complex that contains a subset of trithorax group proteins. *Mol. Cell Biol.* **23**, 140–149 (2003).
- Wysocka, J., Myers, M. P., Laherty, C. D., Eisenman, R. N. & Herr, W. Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. *Genes Dev.* **17**, 896–911 (2003).
- Bellve, A. R. *et al.* Spermatogenic cells of the prepuberal mouse. Isolation and morphological characterization. *J. Cell Biol.* **74**, 68–85 (1977).
- Tachibana, M., Sugimoto, K., Fukushima, T. & Shinkai, Y. Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. *J. Biol. Chem.* **276**, 25309–25317 (2001).
- Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
- O'Carroll, D. *et al.* Isolation and characterization of Suv39h2, a second histone H3 methyltransferase gene that displays testis-specific expression. *Mol. Cell Biol.* **20**, 9423–9433 (2000).
- Miller, T. *et al.* COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proc. Natl Acad. Sci. USA* **98**, 12902–12907 (2001).
- Rogeev, A. *et al.* The *Saccharomyces cerevisiae* Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *EMBO J.* **20**, 7137–7148 (2001).
- Watanabe, D., Sawada, K., Koshimizu, U., Kagawa, T. & Nishimune, Y. Characterization of male meiotic germ cell-specific antigen (Meg 1) by monoclonal antibody TRA 369 in mice. *Mol. Reprod. Dev.* **33**, 307–312 (1992).
- Mahadevaiah, S. K. *et al.* Recombinational DNA double-strand breaks in mice precede synapsis. *Nature Genet.* **27**, 271–276 (2001).
- Pittman, D. L. *et al.* Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for Dmc1, a germline-specific RecA homolog. *Mol. Cell* **1**, 697–705 (1998).
- Yoshida, K. *et al.* The mouse RecA-like gene Dmc1 is required for homologous chromosome synapsis during meiosis. *Mol. Cell* **1**, 707–718 (1998).
- Baudat, F., Manova, K., Yuen, J. P., Jasin, M. & Keeney, S. Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Mol. Cell* **6**, 989–998 (2000).
- Romanienko, P. J. & Camerini-Otero, R. D. The mouse *Spo11* gene is required for meiotic chromosome synapsis. *Mol. Cell* **6**, 975–987 (2000).
- Turner, J. M. *et al.* Silencing of unsynapsed meiotic chromosomes in the mouse. *Nature Genet.* **37**, 41–47 (2005).
- Baarends, W. M. *et al.* Silencing of unpaired chromatin and histone H2A ubiquitination in mammalian meiosis. *Mol. Cell Biol.* **25**, 1041–1053 (2005).
- Dutta, R. & Inouye, M. GHKL, an emergent ATPase/kinase superfamily. *Trends Biochem. Sci.* **25**, 24–28 (2000).
- Peters, A. H. *et al.* Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* **107**, 323–337 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Nakatsuji and S. Chuma for anti-SCP3 antibody; Y. Nishimune for TRA369 antibody; T. Noce for anti-mVASA antibody; H. Kai and Y. Seki for pcDNA3GAL4DBD; J. Miyazaki for pCAGGS; M. Saitou for confocal microscopy; and M. Tachibana and Y. Shinkai for GST-fused histone H3 expression vectors. This work was supported in part by CREST of JST (Japan Science and Technology Agency), and by grants-in-aid and Special Coordinating Funds for Promoting Science and Technology from the Ministry of Education, Science, Sports and Culture of Japan.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.M. (y matsui@idac.tohoku.ac.jp).

Chromatin remodelling at a DNA double-strand break site in *Saccharomyces cerevisiae*

Toyoko Tsukuda¹, Alastair B. Fleming¹, Jac A. Nickoloff¹ & Mary Ann Osley¹

The repair of DNA double-strand breaks (DSBs) is crucial for maintaining genome stability. Eukaryotic cells repair DSBs by both non-homologous end joining and homologous recombination. How chromatin structure is altered in response to DSBs and how such alterations influence DSB repair processes are important issues. In vertebrates, phosphorylation of the histone variant H2A.X occurs rapidly after DSB formation¹, spreads over megabase chromatin domains, and is required for stable accumulation of repair proteins at damage foci². In *Saccharomyces cerevisiae*, phosphorylation of the two principal H2A species is also signalled by DSB formation, which spreads ~40 kb in either direction from the DSB³. Here we show that near a DSB phosphorylation of H2A is followed by loss of histones H2B and H3 and increased sensitivity of chromatin to digestion by micrococcal nuclease; however, phosphorylation of H2A and nucleosome loss occur independently. The DNA damage sensor MRX⁴ is required for histone loss, which also depends on INO80, a nucleosome remodelling complex⁵. The repair protein Rad51 (ref. 6) shows delayed recruitment to DSBs in the absence of histone loss, suggesting that MRX-dependent nucleosome remodelling regulates the accessibility of factors directly involved in DNA repair by homologous recombination. Thus, MRX may regulate two pathways of chromatin changes: nucleosome displacement for efficient recruitment of homologous recombination proteins; and phosphorylation of H2A, which modulates checkpoint responses to DNA damage².

To elucidate the chromatin pathways leading to DSB repair in *S. cerevisiae*, here we have used a *MAT α* haploid strain that lacks *HMR* and *HML* donor sequences and carries a galactose-inducible homothallic switching endonuclease gene (*HO*)⁷. In this strain, HO endonuclease introduces a DSB at *MAT* that can be repaired only by non-homologous end joining, although the principal proteins involved in homologous recombination are recruited to the break site⁶. We analysed chromatin structure over a region of 12–20 kb encompassing the DSB by chromatin immunoprecipitation (ChIP) followed by real-time polymerase chain reaction (PCR), which provides a sensitive measurement of the kinetics and spatial distribution of chromatin changes and the recruitment of repair proteins around the break site.

Budding yeast H2A is phosphorylated on Ser 129 by the ATM and ATR homologues Tel1 and Mec1, respectively⁸. In agreement with a previous report³, we found that phosphorylated H2A (γ -H2A) accumulated rapidly and extensively on either side of the DSB, and that the amount of γ -H2A was less near the DSB than at a position 6-kb distant (Fig. 1a and Supplementary Fig. 3a). This latter result suggested that nucleosome integrity is lost near the DSB. The nucleosome consists of 146 bp of DNA wrapped about twice around a histone octamer comprising a (H3–H4)₂ tetramer and two H2A–H2B dimers. To determine whether nucleosome stability was altered at the DSB, we did ChIP in strains expressing either Flag–H2B or

Flag–H3. The quantities of both histones decreased 60–90 min after HO induction and were reduced threefold by 120 min (Fig. 1a). The comparable loss of both histones suggests that whole nucleosomes were displaced from chromatin near the DSB. Although no histone loss occurred in the first 30 min of DSB induction, quantities of γ -H2A were about fourfold less near the DSB than at distal sites at this time (Fig. 1a). This difference may reflect phosphatase activity near the DSB or enhanced phosphorylation at distal sites.

To confirm that nucleosomes were remodelled near the DSB, we analysed the sensitivity of *MAT α* chromatin to micrococcal nuclease (MNase; Fig. 1b). Before DSB induction, a strong MNase ladder reflects positioned nucleosomes⁹. We found that after DSB formation the nucleosome ladder became progressively less organized with time (Fig. 1b and Supplementary Figs 1 and 2). The alteration in the nucleosome pattern closely paralleled histone depletion, indicating that nucleosome integrity is compromised around the DSB.

Histone eviction during transcription often depends on histone-modifying or ATP-dependent nucleosome remodelling factors^{10–12}. Because the accumulation of γ -H2A temporally preceded histone loss, we first examined its role in nucleosome displacement by eliminating the MRX (Mre11–Rad50–Xrs2) complex. Phosphorylation of H2A.X by ATM in vertebrates requires the homologous MRN (Mre11–Rad50–Nbs1) complex¹³. MRX is one of the earliest factors recruited to yeast DSBs, and it regulates the ATM homologue Tel1 (ref. 14). In *mre11 Δ* , γ -H2A was not abolished at the DSB, although its overall amounts were reduced 2–3-fold (Supplementary Fig. 3c). Histone loss was significantly impeded (Fig. 2a), however, and Flag–H2B was present 3 h after DSB induction (data not shown). In addition, the nucleosome ladder at *MAT α* did not change over this time period, indicating that the chromatin structure remained intact (Fig. 2b and Supplementary Figs 1 and 2).

Because these data suggested that phosphorylation of H2A and histone loss might not be directly coupled, we examined histone occupancy in an H2A mutant lacking the Ser 129 phosphorylation site (*hta1/hta2-S129**; Fig. 2a). Although γ -H2A could not be detected at the DSB in this mutant (data not shown), Flag–H3 was lost to the same extent as in a wild-type strain. Thus, histone eviction depends on MRX but not on γ -H2A.

Next, we determined whether an ATP-dependent nucleosome remodelling activity was required for histone loss. We focused on INO80 (a complex named for the Ino80 protein) because an *ino80 Δ* mutant shows sensitivity to agents that cause DSBs, INO80 can move nucleosomes *in vitro*^{5,15}, and the Ino80 protein is enriched at *MAT* after DSB induction^{16–18}. Because the *ino80 Δ* mutation causes inviability in our strain background, we used an *arp8 Δ* mutant, which is defective in the INO80 ATPase and chromatin remodelling activities and is sensitive to DNA damage¹⁵. In *arp8 Δ* , Flag–H3 and Flag–H2B were retained at the DSB for 2 h, and were evicted only 3 h after break formation (Fig. 3a and Supplementary Fig. 3b). Consistent with postponed histone eviction, there was a similar delay in nucleosome

¹Department of Molecular Genetics and Microbiology, University of New Mexico School of Medicine, 915 Camino de Salud Albuquerque, New Mexico 87131, USA.

disruption (Fig. 3c and Supplementary Figs 1 and 2). Thus, *arp8Δ* significantly slows but does not eliminate histone loss. Together, these results indicate that INO80-dependent chromatin remodelling increases the kinetics of nucleosome displacement at a DSB and that this activity depends on MRX but not γ -H2A.

In wild-type cells, histone loss begins ~ 60 min after DSB induction, which temporally coincides with the 5' to 3' resection of DNA ends¹⁹. Moreover, an *mre11Δ* mutant, which is defective in histone

eviction, is also defective in end processing⁷. We tested whether the slower rate of histone eviction in *arp8Δ* results from a slower rate of DSB formation or end processing, and found that both cleavage and resection were similar in wild-type and *arp8Δ* (Supplementary Fig. 4), in contrast to a previous report¹⁷. This finding also indicates that nucleosomes can associate *in vivo* with single-stranded DNA, as observed *in vitro*²⁰, and that nucleases can resect DNA that is nucleosomal. Thus, single-stranded DNA is necessary but not sufficient for histone eviction, and the INO80 ATPase is required for chromatin remodelling at the DSB.

Because Ino80 physically associates with γ -H2A and γ -H2A recruits Ino80 to the *MAT* DSB^{16–18}, our finding of γ -H2A-independent histone displacement was puzzling. To address this issue, we measured Ino80 association with *MAT* α before DSB induction and found that it was present even in the absence of the break (Fig. 3b, –Gal). As shown by DNA microarray studies, this pool of Ino80 probably contributes to *MAT* α transcription²¹, and we confirmed that *MAT* $\alpha 1$ and *MAT* $\alpha 2$ transcript levels are 2–3-fold lower in *arp8Δ*

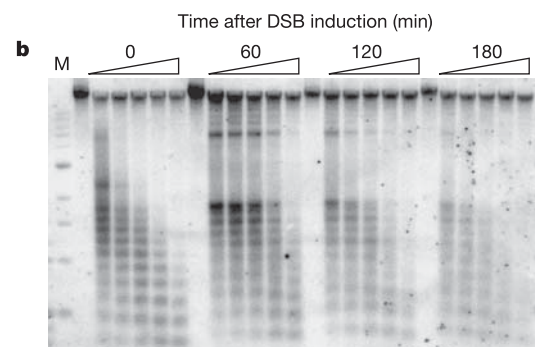
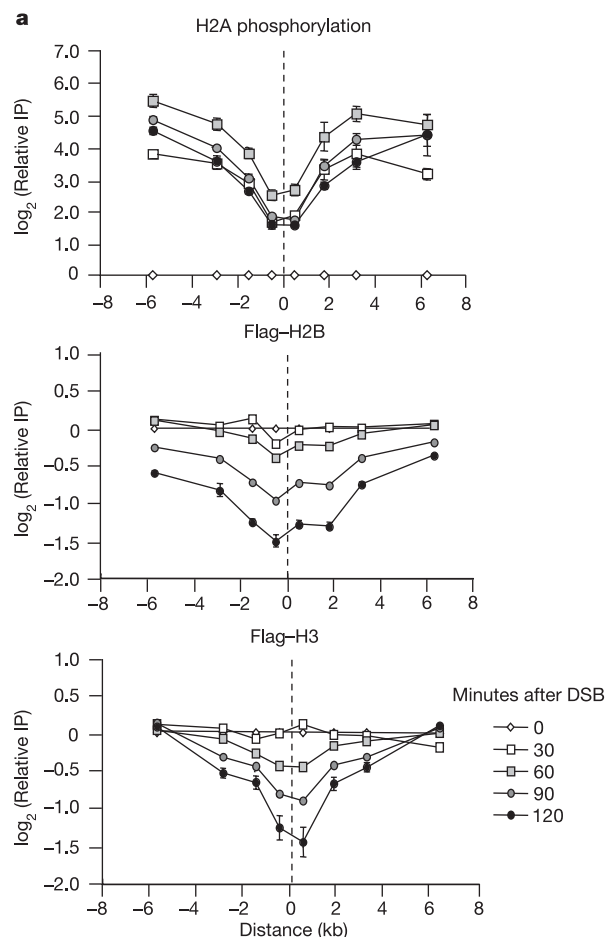


Figure 1 | Chromatin changes at the *MAT* α DSB. **a**, A DSB was induced at *MAT* α in strains expressing Flag–H2B or Flag–H3, and ChIP was done with antibodies against γ -H2A or Flag. DNA was analysed by real-time PCR using primers corresponding to sequences on the left (–) or the right (+) side of the DSB (0), and results were normalized to the ratio of immunoprecipitation (IP) to input DNA at time 0. Data are the mean $\pm$ s.e.m. **b**, Nuclei were prepared after DSB induction, and chromatin was digested with MNase and subjected to Southern blot analysis using a probe for *MAT* α DNA. The triangles denote increasing times of MNase digestion. M indicates a 1-kb DNA ladder.

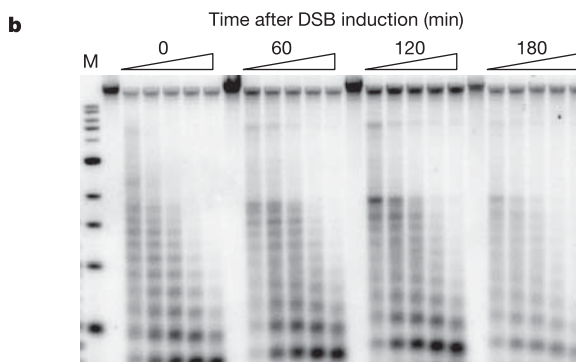
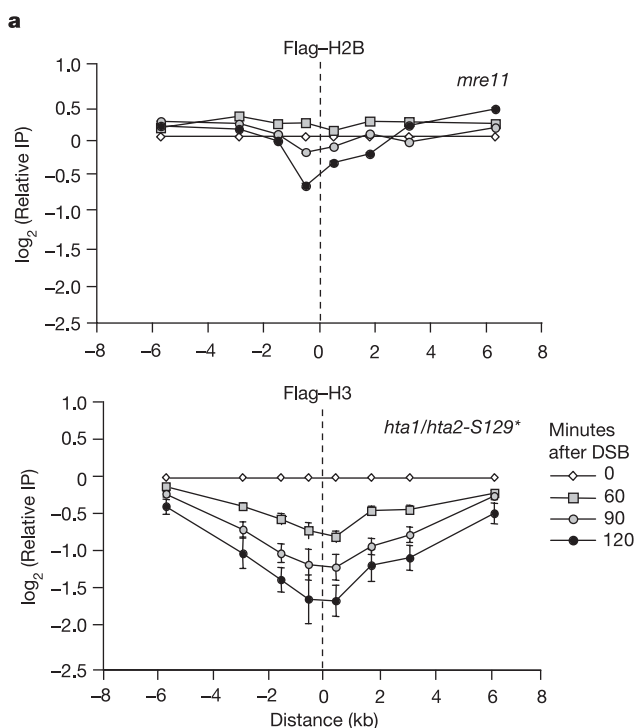


Figure 2 | MRX is involved in histone loss at the *MAT* α DSB. **a**, A DSB was induced at *MAT* α , and ChIP was done with antibodies against Flag in an *mre11::Kan-MX* strain expressing Flag–H2B and an *hta1/hta2-S129** strain expressing Flag–H3. DNA was analysed by real-time PCR on both the left (–) and the right (+) side of the DSB (0). Data are the mean $\pm$ s.e.m. **b**, MNase analysis was done on nuclei isolated from the *mre11::Kan-MX* strain by Southern blot analysis as described in Fig. 1b.

(data not shown). After DSB induction (Fig. 3b, +Gal), Ino80 preferentially accumulated on the right side of the DSB and, in contrast to the pre-existing pool of Ino80, this association required γ -H2A^{16,17} (Supplementary Fig. 5). In addition, 2 h after break induction ~50% of the pre-existing Ino80 pool was lost at a point 0.5 kb from the left side of the DSB, a region corresponding to the *MAT* α promoter (Fig. 3b, +Gal). Together, these data suggest that a pre-existing γ -H2A-independent pool of Ino80 related to *MAT* α transcription is responsible for histone displacement, whereas a newly recruited pool performs another function. Ino80 binds to both DNA and histones^{5,15}, and notably DNA sequences to the right of the *MAT* DSB initially participate in strand invasion during homologous recombination, suggesting that the newly recruited Ino80 has a role in this process.

The INO80 complex is associated with the repair of DSBs by homologous recombination¹⁷ (Supplementary Fig. 6), indicating that histone depletion might affect the recruitment of homologous recombination proteins to the *MAT* DSB. HO-generated DSBs are stable for almost 60 min (ref. 19), and the recruitment of homologous recombination proteins such as Rad51 and Rad54 is delayed until the broken ends are processed to 3' single-stranded tails⁶. Replication protein A (RPA) binds first, spreads along the DNA and is subsequently displaced by Rad51 (refs 22, 23). We found that RPA was efficiently recruited to the DSB in both wild-type and *arp8* Δ cells, with slightly faster kinetics in *arp8* Δ (Fig. 4a). In both strains, RPA eventually spread 6.3 kb from the DSB, confirming that strand resection occurs normally in *arp8* Δ . Recruitment and spreading of

Rad51 were delayed in *arp8* Δ (Fig. 4a), however, and correlated precisely with the delay in histone eviction in this mutant. In *mre11* Δ there was a pronounced delay in the binding of both RPA and Rad51 (Fig. 4a), consistent with the delayed resection of DNA ends^{7,24} (data not shown) and extended period of nucleosome retention in this strain. In marked contrast, recruitment of Rad51 in the *hta1/hta2-S129** mutant was indistinguishable from that in wild type (data not shown), supporting the observation that histone displacement proceeds with normal timing in the absence of γ -H2A. Together, these data suggest that nucleosome eviction controls the rate at which Rad51 displaces RPA from resected DNA during repair by homologous recombination, and that this displacement depends on MRX and the INO80 complex.

The above results indicate that a temporal sequence of events alters chromatin at DSBs. γ -H2A first spreads over a large domain around the break. Nucleosomes are then displaced near the DSB through the remodelling activity of the INO80 complex, which increases the kinetics of histone loss. Nucleosome loss at *MAT* requires MRX, one of the earliest factors to be recruited to DSBs *in vivo*. We propose that MRX regulates nucleosome displacement through its role in two pathways (Fig. 4b). The first pathway involves the MRX-dependent resection of DNA ends. When strand resection is prevented, or significantly delayed as in *mre11* Δ , nucleosome displacement cannot occur efficiently. Resection itself is not sufficient for nucleosome loss, however, and MRX also functions in a second pathway that controls

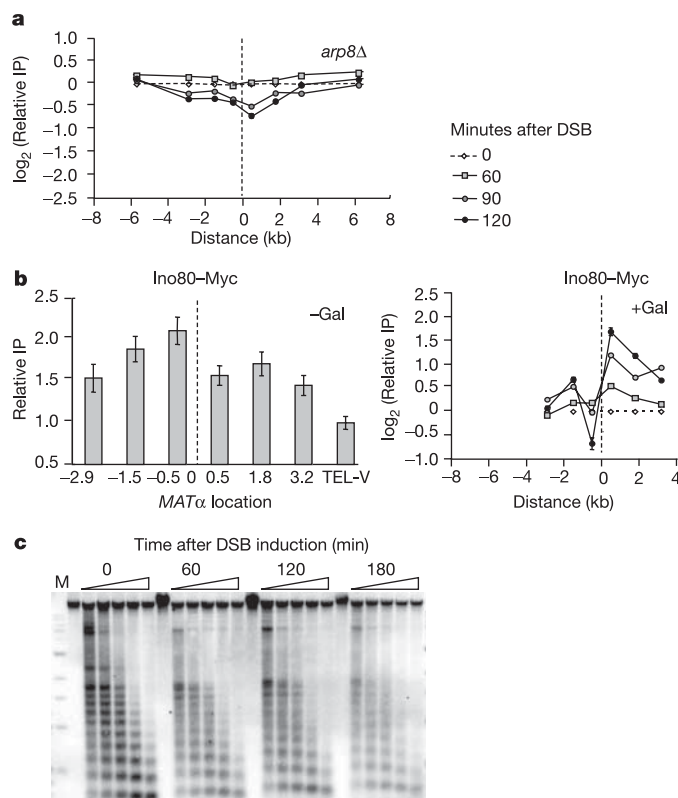


Figure 3 | The INO80 complex is required for histone eviction at the *MAT* α DSB. **a**, A DSB was induced at *MAT* α in an *arp8::Kan-MX* mutant expressing Flag-H3, and ChIP was done with antibodies against Flag. Precipitated DNA was quantified as described in Fig. 1a. **b**, ChIP was done with antibodies against Myc in a wild-type strain that contained Ino80-Myc before (-Gal) and after (+Gal) DSB induction. Ino80-Myc association was normalized to histone H3 occupancy. **c**, MNase digestion was done on nuclei isolated from an *arp8::Kan-MX* mutant as described in Fig. 1b. Data in **a** and **b** are the mean $\pm$ s.e.m.

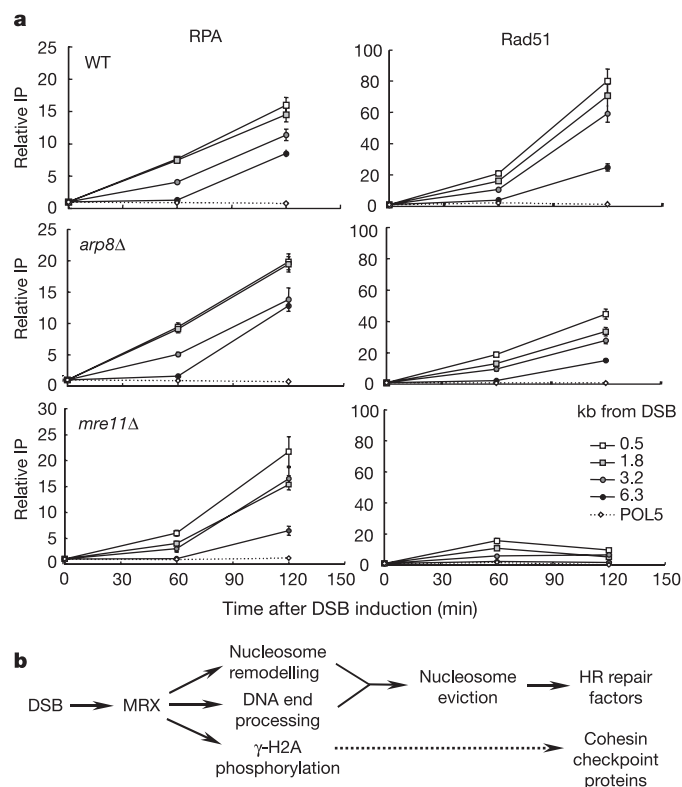


Figure 4 | MRX and INO80 are required for recruiting Rad51 to the *MAT* α DSB. **a**, ChIP was done in wild type, *arp8* Δ or *mre11* Δ with antibodies against RPA (left) or Rad51 (right) after DSB induction at *MAT* α . DNA on the right side of the DSB was analysed by real-time PCR. Data are the mean $\pm$ s.e.m.. **b**, MRX controls chromatin remodelling at DSBs. MRX is recruited to DSBs, where it regulates DNA end processing and Tel1-dependent phosphorylation of H2A. MRX regulates nucleosome remodelling through INO80, leading to nucleosome eviction and the efficient recruitment of proteins involved in homologous recombination. Phosphorylation of H2A is independent of nucleosome displacement and controls the accumulation of checkpoint proteins, as well as cohesin, at DSBs.

INO80-dependent nucleosome remodelling. These two MRX-dependent pathways converge to promote nucleosome eviction. This convergence of two pathways could account for the more severe defect in nucleosome displacement observed in *mre11Δ* as compared with *arp8Δ*, although we cannot rule out the possibility that another nucleosome remodelling factor shares a redundant role with INO80 (ref. 25).

Histone displacement occurs in a region of the *MATα* locus that includes the promoter and $\alpha 1$ and $\alpha 2$ coding regions, raising the possibility that transcription itself is involved in nucleosome loss through INO80-dependent chromatin remodelling. At highly transcribed genes, an increase in RNA polymerase II correlates with a decrease in histone occupancy^{11,12}. Our preliminary data indicate that *MATα1* transcripts transiently accumulate on DSB induction, coinciding with an increase in RNA polymerase II and a loss of nucleosomes at the $\alpha 1$ TATA site; these effects are significantly reduced in *arp8Δ* (T.T. and A.B.F., unpublished data). Thus, the DSB might signal through MRX to the *MAT* transcription machinery, which would then use the remodelling activity of INO80 to displace histones. This pathway might reflect a mechanism that couples transcription-dependent chromatin remodelling to DSB repair in active promoter regions.

INO80-promoted nucleosome displacement regulates the recruitment of proteins with direct roles in homologous recombination. Rad51 does not efficiently replace RPA in the absence of nucleosome loss, probably owing to the delayed recruitment of Rad52 (T.T., unpublished data), which cooperates with Rad51 to displace RPA^{6,26}. In agreement with the idea that INO80 has a role in repair by homologous recombination, non-homologous end joining is relatively unaffected in *arp8Δ*, whereas *arp8Δ* and homologous recombination mutants show similar sensitivities to DNA damaging agents¹⁷ (Supplementary Fig. 6) and *arp8Δ* shows a 2–3-fold decrease in allelic recombination (S. Krishna and J.A.N., unpublished data). In addition to nucleosome displacement, the cellular response to DNA damage also requires the formation of γ -H2A. Both forms of chromatin remodelling are ultimately dependent on MRX but are independent of one another, comprising two parallel pathways of chromatin change that have different but complementary roles in the DSB repair response (Fig. 4b). MRX-dependent nucleosome remodelling by INO80 is required for efficient recruitment of homologous recombination proteins, whereas MRX-dependent H2A phosphorylation through Tel1 mediates the accumulation of cohesin and checkpoint proteins at DSB sites^{14,24,27}.

METHODS

ChIP and PCR analysis. We fixed 50 ml of cells with 1% formaldehyde for either 60 min or 15 min (histones) and lysed them in FA buffer²⁸. Immunoprecipitation was done with 5 absorbance units at 600 nm ($A_{600\text{ nm}}$ units) for histones and 20 $A_{600\text{ nm}}$ units for all other proteins. We used 5 μ l of antibodies against Myc (9E10, Upstate Biotechnology), 45 μ l of agarose beads conjugated to antibodies against Flag (M2, Sigma), 5 μ l of antibodies against H2A.X phosphorylated on Ser 139 (Upstate Biotechnology), and 2 μ l or 4 μ l of antibodies against RPA or Rad51 (a gift from W.-D. Heyer). Extracted DNA was analysed by real-time PCR using a SYBR Green master mix (ABI) in an 7000 sequence detection system (ABI). DNA primers were designed to span a region of 12–20 kb around the HO cut site at *MATα*, as well as a region of the *POL5* gene; primer sequences are available from the authors on request.

Dissociation curve analysis of the amplified DNA melting temperature showed that the each primer set gave a single and specific product. The immunoprecipitation data were normalized to the *POL5* gene, where DSB does not occur, to correct for experimental variation and loss of DNA at the DSB site⁶. Most experiments were repeated at least twice and, in each experiment, PCR reactions were done in triplicate. The relative immunoprecipitation value represents the ratio of immunoprecipitated DNA to *POL5* input DNA after HO induction normalized to the ratio of immunoprecipitated DNA to *POL5* input DNA before HO induction. Standard errors were calculated by using error analysis for more than one measurement²⁹. This analysis finds the largest error source by combining standard errors of two measurements, in this case, measurements of immunoprecipitated DNA and input DNA.

Micrococcal nuclease digestion. We collected 1-litre cultures just before and at 60-min intervals after DSB induction. Nuclei were prepared under conditions maintaining HO induction (plus 2% galactose), and 1.5 units of MNase (Worthington) were added to 0.25 ml of nuclei as described³⁰. Samples were removed at 1-, 2-, 4-, 8- and 16-min intervals, and DNA was purified and separated by electrophoresis on a 1.25% agarose-TBE gel. DNA was blotted to Genescreen (Dupont-NEN) and hybridized to an 800-bp *MATα* fragment that was ~200 bp from the right side of the HO-induced DSB. Radioactive images were captured on a Model 860 Phosphorimager (Storm) and band intensities were quantified using ImageQuant TL software (Molecular Dynamics).

Received 23 November 2004; accepted 22 August 2005.

- Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J. Biol. Chem.* **273**, 5858–5868 (1998).
- Arkady, C. *et al.* Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks. *Nature Cell Biol.* **5**, 675–679 (2003).
- Shroff, R. *et al.* Distribution and dynamics of chromatin modification induced by a defined DNA double-strand break. *Curr. Biol.* **14**, 1703–1711 (2004).
- Usui, T., Ogawa, H. & Petrini, J. H. A DNA damage response pathway controlled by Tel1 and the Mre11 complex. *Mol. Cell* **7**, 1255–1266 (2001).
- Shen, X., Mizuguchi, G., Hamiche, A. & Wu, C. A chromatin remodelling complex involved in transcription and DNA processing. *Nature* **406**, 541–544 (2000).
- Sugawara, N., Wang, X. & Haber, J. E. *In vivo* roles of Rad52, Rad54, and Rad55 proteins in Rad51-mediated recombination. *Mol. Cell* **12**, 209–219 (2003).
- Lee, S. E. *et al.* *Saccharomyces* Ku70, Mre11/Rad50 and RPA proteins regulate adaptation to G2/M arrest after DNA damage. *Cell* **94**, 399–409 (1998).
- Downs, J. A., Lowndes, N. F. & Jackson, S. P. A role for *Saccharomyces cerevisiae* histone H2A in DNA repair. *Nature* **408**, 1001–1004 (2000).
- Weiss, K. & Simpson, R. T. High-resolution structural analysis of chromatin at specific loci: *Saccharomyces cerevisiae* silent mating type locus *HMLα*. *Mol. Cell. Biol.* **18**, 5392–5403 (1998).
- Reinke, H. & Horz, W. Histones are first hyperacetylated and then lose contact with the activated *PHO5* promoter. *Mol. Cell* **11**, 1599–1607 (2003).
- Kristjuhan, A. & Svestrup, J. Q. Evidence for distinct mechanisms facilitating transcript elongation through chromatin *in vivo*. *EMBO J.* **23**, 4243–4252 (2004).
- Schwabish, M. A. & Struhl, K. Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. *Mol. Cell. Biol.* **24**, 10111–10117 (2004).
- Uziel, T. *et al.* Requirement of the MRN complex for ATM activation by DNA damage. *EMBO J.* **22**, 5612–5621 (2003).
- Lisby, M., Barlow, J. H., Burgess, R. C. & Rothstein, R. Choreography of the DNA damage response: spatiotemporal relationships among checkpoint and repair proteins. *Cell* **118**, 699–713 (2004).
- Shen, X., Ranallo, R., Choi, E. & Wu, C. Involvement of actin-related proteins in ATP-dependent chromatin remodeling. *Mol. Cell* **12**, 147–155 (2003).
- Morrison, A. J. *et al.* INO80 and γ -H2AX interaction links ATP-dependent chromatin remodeling to DNA damage repair. *Cell* **119**, 767–775 (2004).
- van Attikum, H., Fritsch, O., Hohn, B. & Gasser, S. M. Recruitment of the INO80 complex by H2A phosphorylation links ATP-dependent chromatin remodeling with DNA double-strand break repair. *Cell* **119**, 777–788 (2004).
- Downs, J. A. *et al.* Binding of chromatin-modifying activities to phosphorylated histone H2A at DNA damage sites. *Mol. Cell* **16**, 979–990 (2004).
- Frank-Vaillant, M. & Marcand, S. Transient stability of DNA ends allows nonhomologous end joining to precede homologous recombination. *Mol. Cell* **10**, 1189–1199 (2002).
- Palter, K. B., Foe, V. E. & Alberts, B. M. Evidence for the formation of nucleosome-like histone complexes on single-stranded DNA. *Cell* **18**, 451–467 (1979).
- Mizuguchi, G. *et al.* ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303**, 343–348 (2004).
- Wang, X. & Haber, J. E. Role of *Saccharomyces* single-stranded DNA-binding protein RPA in the strand invasion step of double-strand break repair. *PLoS Biol.* **2**, 0104–0112 (2004).
- Kantake, N., Sugiyama, T., Kolodner, R. D. & Kowalczykowski, S. C. The recombination-deficient mutant RPA (*rfal-117*) is displaced slowly from single-stranded DNA by Rad51 protein. *J. Biol. Chem.* **278**, 23410–23417 (2003).
- Unal, E. *et al.* DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain. *Mol. Cell* **16**, 991–1002 (2004).
- Chai, B., Huang, J., Cairns, B. & Laurent, B. C. Distinct roles for the Rsc and Swi/Snf ATP-dependent chromatin remodelers in DNA double-strand break repair. *Genes Dev.* **19**, 1656–1661 (2005).
- Miyazaki, T., Bressan, D. A., Shinohara, M., Haber, J. E. & Shinohara, A. *In vivo* assembly and disassembly of Rad51 and Rad52 complexes during double-strand break repair. *EMBO J.* **23**, 939–949 (2004).
- Nakamura, T. M., Du, L. L., Redon, C. & Russell, P. Histone H2A phosphorylation controls Crb2 recruitment at DNA breaks, maintains checkpoint arrest, and influences DNA repair in fission yeast. *Mol. Cell. Biol.* **24**, 6215–6230 (2004).

28. Kuo, M. H. & Allis, C. D. *In vivo* cross-linking and immunoprecipitation for studying dynamic protein:DNA associations in a chromatin environment. *Methods* **19**, 425–433 (1999).
29. Lichten, W. *Data and Error Analysis in the Introductory Physics Laboratory* (Allyn and Bacon, Newton, MA, 1988).
30. Fleming, A. B. & Pennings, S. Antagonistic remodelling by Swi-Snf and Tup1-Ssn6 of an extensive chromatin region forms the background for *FLO1* gene regulation. *EMBO J.* **20**, 5219–5231 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Haber, X. Shen, J. Downs, M. Christman and V. Zakian for strains or plasmids; W.-D. Heyer for antibodies against RPA and Rad51; C. Hillyer and N. Clark for technical assistance; and C.-F. Kao for comments. This work was supported by grants from the NIH. (to M.A.O. and to J.A.N.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.A.O. (mosley@salud.unm.edu).

LETTERS

Force production by disassembling microtubules

Ekaterina L. Grishchuk^{1,2}, Maxim I. Molodtsov^{1,3}, Fazly I. Ataullakhanov^{3,4,5} & J. Richard McIntosh¹

Microtubules (MTs) are important components of the eukaryotic cytoskeleton: they contribute to cell shape and movement, as well as to the motions of organelles including mitotic chromosomes. MTs bind motor enzymes that drive many such movements, but MT dynamics can also contribute to organelle motility^{1–8}. Each MT polymer is a store of chemical energy that can be used to do mechanical work, but how this energy is converted to motility remains unknown. Here we show, by conjugating glass microbeads to tubulin polymers through strong inert linkages, such as biotin–avidin, that depolymerizing MTs exert a brief tug on the beads, as measured with laser tweezers. Analysis of these interactions with a molecular-mechanical model of MT structure and force production^{9,10} shows that a single depolymerizing MT can generate about ten times the force that is developed by a motor enzyme; thus, this mechanism might be the primary driving force for chromosome motion. Because even the simple coupler used here slows MT disassembly, physiological couplers may modulate MT dynamics *in vivo*.

It has previously been proposed that MTs might generate force through their unusual mechano-chemistry^{5,10}. These polymers assemble from tubulin dimers that bind GTP and hydrolyse it after polymerization. The GDP-bound dimer conformation is bent relative to its GTP-bound counterpart^{11,12}, but this bend is constrained by MT geometry so that some energy from GTP hydrolysis is stored in the polymer lattice^{13,14}. During depolymerization, the accumulated stress forces the strands of GDP-bound dimers, or ‘protofilaments’ (PFs), to arch out in a ‘ram’s horn’ configuration¹⁵. Thus, each MT is a reservoir of chemical energy that can be harnessed to do mechanical work. Two aspects of this mechanism have gone uncharacterized: the power-stroke, which is probably generated by curving PFs; and the coupling devices that might convert this energy into motility. Here we have conjugated glass microbeads to MTs and used laser tweezers to measure the forces exerted on the beads as the MTs depolymerize.

Biotinylated MTs were grown from *Tetrahymena* ‘pellicles’ (that is, lysed, de-ciliated protozoa), which will bind to the coverslip that tops a microscope chamber. Such pellicles define the orientation of the MTs that grow from them (the fast-growing or ‘plus’ ends are distal), and they provide a firm anchorage for the minus ends⁶ (Fig. 1a). These labile MTs were capped with rhodamine-conjugated tubulin without biotin by extending existing polymers in the presence of GMPCPP, a slowly hydrolysable GTP analogue that forms stable MTs¹⁶. Glass microbeads coated with streptavidin were then allowed to bind to the biotinylated MT segments (Fig. 1b). To begin an experiment, the stable, rhodamine-labelled MT caps were dispersed by green light, which excites rhodamine fluorescence and induces the caps to break up¹⁷. The labile segments of the MTs were then exposed and depolymerized from their plus ends.

As the MTs depolymerized, the bound beads were released. We used laser tweezers to investigate the accompanying mechanical transients. Before a typical experiment our tweezers pulled an

MT-attached bead toward the plus end of the polymer with an average of 0.5 pN. The position of the bead relative to the centre of the trap was measured with the much weaker beam from a tracking laser, imaged on a quadrant photodiode (QPD)^{18,19} (Fig. 2). Non-random bead motions were usually detected within 1 min of the green flash. Sometimes these motions were complex and lasted for 5–40 s, but ultimately the bead always returned to the centre of the trapping beam. Thereafter, it showed the increased brownian movement characteristic of a free bead (Fig. 2b). We interpreted these complex motions as the result of the bead’s being attached to several MTs that depolymerized independently at slightly different times, causing a series of changes in bead position. Because we could not know how many MTs were initially attached to a bead, we focused our analysis on the last motion-inducing event, which resulted from depolymerization of the last bead-associated MT. Bead release was confirmed by moving the stage to assure that the bead remained at the centre of the trap.

In 61 out of 144 experiments, the bead moved slightly away from the centre of the trap before relaxing to the free position (Fig. 2c–e). Because the bead was monitored by the QPD in three dimensions, we could compare the direction of its movements with the orientation of the MT, as observed with differential interference contrast optics. Bead movement was directed towards the minus end in 92% of the

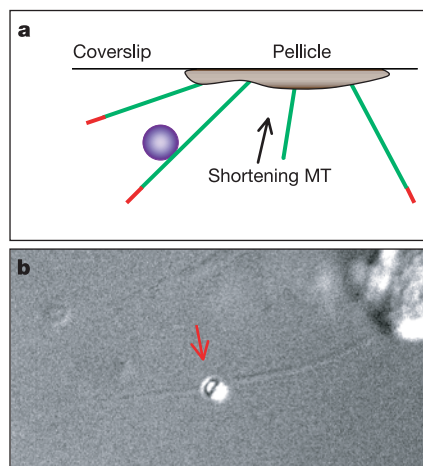


Figure 1 | Experimental design. **a**, The experimental system (not to scale). Photodamage of the plus-end, rhodamine-conjugated tubulin caps (red) exposes dynamic MTs (green). As the MTs depolymerize, bead motions are measured with laser tweezers. **b**, Differential interference contrast image of a 1-μm bead (arrow) attached to an apparently single MT. Although it is not possible to count the number of MTs attached to each bead at the beginning of an experiment, the narrowness of the histogram of the force magnitudes that developed (Fig. 3a) suggests that almost all of our observations derive from pulls generated by single MTs.

¹MCD Biology Department, University of Colorado at Boulder, Colorado 80309-0347, USA. ²Institute of General Pathology and Pathophysiology, Moscow 125315, Russia.

³National Research Centre for Haematology, Moscow 125167, Russia. ⁴Physics Department, Moscow State University, Moscow 119992, Russia. ⁵Institute of Theoretical and Experimental Biophysics, Pushchino 142292, Russia.

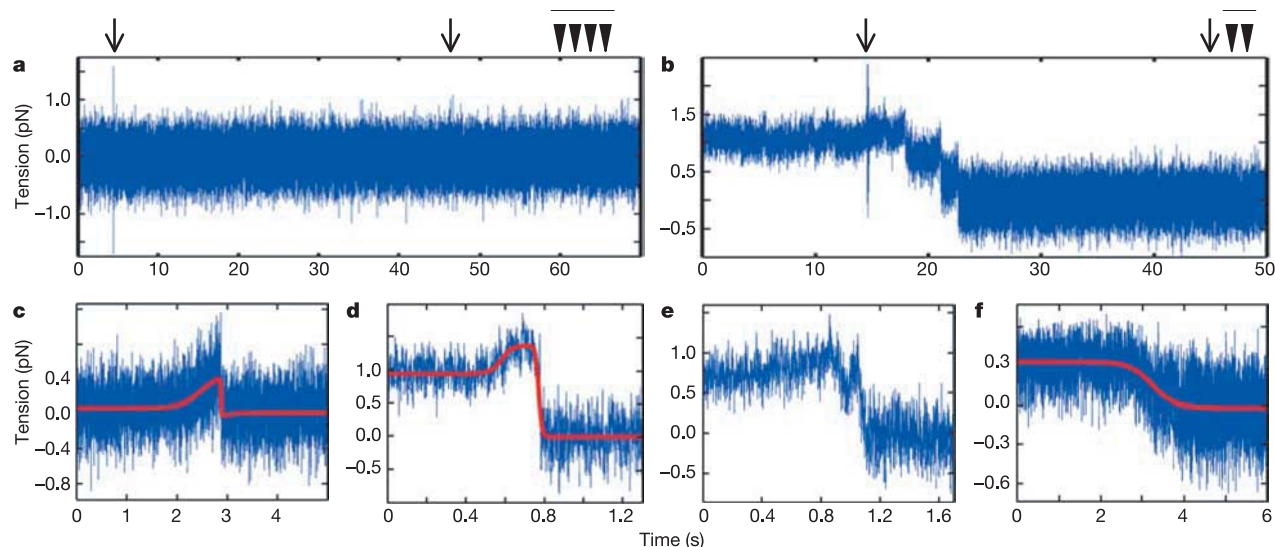


Figure 2 | Example signals. Unprocessed QPD data show changes in bead position and thus force versus time. **a**, A trapped bead not bound to an MT has undergone all routine manipulations: illumination (arrows mark shutter opening and closing) and stage movements (arrowheads). Only brownian motion is seen. **b**, Brownian movement of an MT-associated bead under tension is reduced relative to that in **a**. Low frequency changes before

illumination are from tiny drifts of the stage or pellicle. Thereafter, several changes in bead position are distinguished before detachment. **c**–**f**, Examples of final events. Red line shows the curve that was fit to the data with equations (1) and (2) in the Supplementary Information. **c**, **d**, Typical force signals. **e**, A rare force signal superimposed on relaxation. **f**, No force signal and only slow relaxation.

experiments, and force in the orthogonal direction was usually $<20\%$ of the paraxial force. Similar experiments were conducted with beads pre-incubated with MT-associated proteins (MAPs) enriched for Tau and microtubule-associated protein 2. These proteins provided a different static MT attachment, but the beads showed depolymerization-dependent, minus-end-directed movements with a frequency comparable to that seen with biotin-streptavidin (in 10 out of 32 trials). Thus, a disassembling MT plus end can develop a pulse of force on beads attached by two distinct static links.

A histogram of force magnitudes (Fig. 3a) shows a mean value of 0.24 ± 0.02 pN ($\pm$ s.e.m.; $n = 54$). The force that developed was unaffected by the addition of $2 \mu\text{M}$ soluble biotin before the initiation of MT depolymerization: beads were not competed from the MTs, the frequency of force-producing events was unchanged, and the forces generated in 15 experiments had the same mean as those without added biotin ($P > 0.05$, Student's t -test). When MAPs were used to couple beads to the MTs, the mean force exerted was 0.22 ± 0.04 pN, which, within experimental error, is equal to the value obtained with biotin-streptavidin coupling. Force magnitudes were independent of both bead-pellicle distance and applied tension (data not shown). The amplitude distribution drops to almost zero at ~ 0.46 pN, suggesting that this is approximately the maximal force that this system can develop.

Consideration of bead curvature and MT geometry suggests that only 1–2 PFs can attach to a bead simultaneously; thus, bead movement must result from the bending of these linear tubulin arrays. Moreover, the radius of the trapped bead was 500 nm, but the segment of MT that exerted force on the bead was much smaller, probably 20–60 nm (see Supplementary Information). The measured force was thus reduced by a lever arm that is the ratio of these distances (~ 10). The force at the surface of the bead was therefore ~ 5 pN, which is about the same as that developed by a single MT-dependent motor enzyme²⁰. If a depolymerizing MT end were tracked by a well-designed coupling device, such as an encircling ring that could move freely on the surface of the polymer²¹, one would expect all the PFs to act in concert, increasing the depolymerization-dependent force to at least 30–65 pN.

We interpret these observations by a model in which MT

depolymerization is triggered distal to the bead, and a wave of disassembly, mediated by outward-curving PFs, propagates towards the minus end of the MT. When the wave reaches the bead and the attached PFs bend out from the MT axis, they exert a brief force on the trapped bead. Such a force should grow in amplitude as the PFs continue bending away from the straight MT surface, while breaking the lateral interactions that constrain the shape of GDP-bound tubulin¹⁰. Indeed, the maximal observed force was always achieved gradually (Fig. 2c, d). A histogram of the duration of PF bending under load (Fig. 3b) has a broad distribution ranging from 0.03 to 9.5 s, which is likely to reflect, in part, the natural variability in rates of MT depolymerization^{22,23}. We tested this supposition by comparing the duration of observed force transient for MTs whose depolymerization rates were varied by altering the concentration of Mg^{2+} ions. Lower Mg^{2+} concentrations, which decrease the rate of MT depolymerization²³, produce longer force transients (Fig. 3c, filled bars). At physiological concentrations of Mg^{2+} (1–4 mM), the duration of force signal and the estimated number of dimers involved suggest that PFs bend under load at a rate of 8–12 dimers s^{-1} (see Supplementary Information). Free MT plus ends, by contrast, depolymerize at ~ 65 dimers per s per PF^{22–25}. The 5–8-fold lower value in our experiments is probably due to both the tension from the trap, which antagonizes PF bending, and the interference provided by tubulin binding to the surface of the bead. Analogous effects may well explain the decreased depolymerization rate of kinetochore-associated MT^{1,26}.

No force signal was observed in $\sim 45\%$ of our experiments ($n = 77$); the bead simply moved unidirectionally to the centre of the trap (Fig. 2f). The speed of the bead during this relaxation was variable (Fig. 3). Occasionally it moved as though it were completely free (<20 ms to equilibrium), but $\sim 92\%$ of the events took longer. Relaxations preceded by force production were generally quicker than those without developed force (Fig. 3d). The same trend was seen for relaxation time and force amplitude: smaller forces were commonly followed by slower relaxations (Fig. 3e). These observations suggest that force production by an asymmetric coupler might be impeded by the same processes that slow bead relaxation.

We propose that the variabilities in force generation and bead relaxation times derive from the stochastic peeling of individual PFs

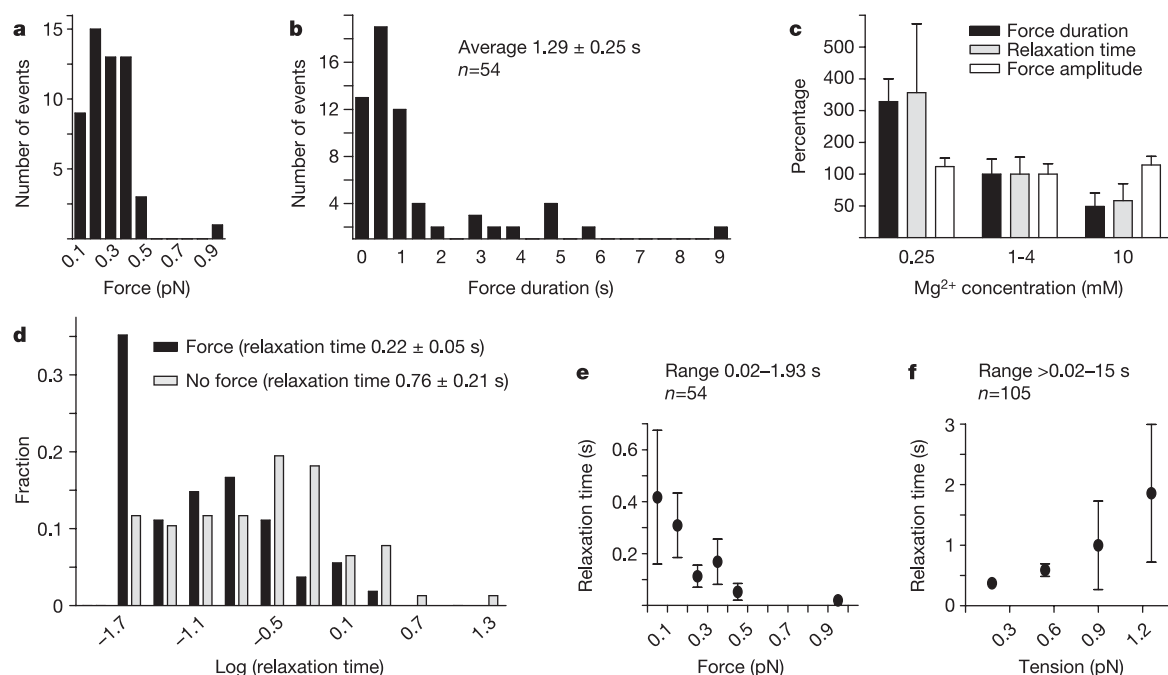


Figure 3 | Analysis of force production. **a**, Histogram of force amplitudes determined from the differences between applied and peak tensions. **b**, Histogram of force duration (that is, the time from the onset of signal increase to the onset of signal decrease). **c**, Force characteristics at different Mg^{2+} concentrations relative to those in physiological conditions (1–4 mM). Both force duration and relaxation time are longer for more slowly depolymerizing MTs. The force amplitudes remain unchanged, however, which suggests that Mg^{2+} ions do not alter the strength of longitudinal

bending forces between dimers. **d**, Histogram of the log of relaxation time (that is, the time required for the signal to decrease from maximum to the free bead value). The ordinate shows the fraction of all experiments in each category. **e**, Relaxation time versus force amplitude in the same signal. **f**, Average relaxation time versus applied tension for slow events (>20 ms). Larger error bars at higher tension result from a smaller number of measurements in this tension range. Error bars represent the s.e.m. with 95% confidence intervals (**c**, **e**, **f**).

(Fig. 4). As lateral bonds between PFs separate, the trapped bead will move differently, depending on which PFs relative to the site of bead attachment are the first to split away from their neighbours and peel outward (see Supplementary Information). For example, the bead-associated PFs could peel away slightly before or simultaneously with the peeling of all other PFs (Fig. 4b); this scenario should generate the strongest force. Alternatively, other PFs might peel before those bound to the bead (Fig. 4c); here, the rigidity of the MT segment downstream from the bead would decrease, and the bead should respond to the applied tension by moving closer to the centre of the trap before the appearance of any force transient generated by the curling of bead-attached PFs. Occasionally force might be generated even while the bead is moving towards the centre of the trap (Fig. 2e). More frequently, however, the superposition of these competing processes would probably cancel the force signal altogether. Thus, force production by the asymmetric coupler might be stochastically attenuated owing to asynchrony in PF splitting.

The slow bead relaxations are striking because they imply that the bead retains its attachment to a partially disassembled MT for up to several seconds. Even within 1 s, splitting of the PFs should progress through 8–65 dimers. Throughout this time, however, the bead is not free; thus, longitudinal bonds between dimers beyond the bead must persist, even if some lateral bonds are lost. (If the longitudinal bonds broke, then the bead would jump to the centre of the trap in <20 ms.) We considered whether increased tension on these PFs might facilitate bond breakage; however, slow relaxation times were frequently associated with stronger, not weaker, applied tensions (Fig. 3f). Thus, the duration of bead attachment to a partially disassembled MT can be enhanced by tension that opposes PF bending.

These force-producing events constitute a ‘single-shot’ mechanism that is mediated by inert, asymmetric couplers. To generate processive movements in association with tubulin depolymerization, there must be a coupling that maintains attachment to the MT

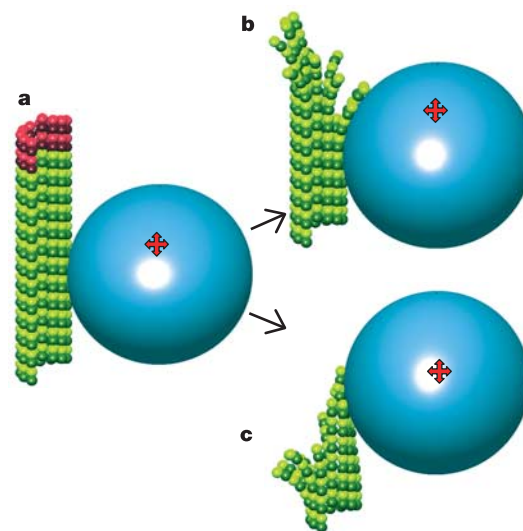


Figure 4 | Models of force production. MT is attached to a bead, which is shown ten times too small. α -Tubulin and β -tubulin are coloured dark and light green, respectively. The red cross marks the trap centre (not shown to scale). **a**, Capped MT with a bead under tension. The bead is attached by roughly three dimers per two adjacent PFs. Red indicates the rhodamine-labelled GMPCPP cap, which was longer and further from the bead than shown. **b**, In scenario 1, the bead-attached PFs peel at the same time or sooner than others. As they curl, the bead is pushed from the trap centre, producing maximal force. **c**, In scenario 2, the bead-attached PFs peel after others, which leaves the bead hanging onto adherent PFs of various lengths. This structure is not rigid and, as it changes owing to continuing disassembly, the tension applied causes the bead to move slowly to the centre of the trap (see Supplementary Information).

surface as the dimers dissociate. Motor enzymes can provide this function^{7,8}. The rings formed by the DAM complexes that bind to yeast kinetochores could also contribute to processive movements^{21,27}. One of the traits of such couplings in mitosis is that MTs stay attached to kinetochores even when tension is high, and increased tension can induce kinetochore MTs to switch into a polymerization state²⁸. We previously proposed that this may occur through rescue of PF bending by constraining properties of the coupler²⁹. Our results now suggest a specific molecular mechanism: opposing tension slows PF bending, which in turn inhibits dimer disassembly downstream from the coupler. *In vivo* this process might promote kinetochore MT rescue and contribute to the accuracy of chromosome segregation without requiring any tension-sensing, regulatory enzymes.

METHODS

Instruments and data acquisition. All observations were made on a Zeiss Axiophot2 adapted for laser tweezers as described³⁰ and modified by the addition of a tracking laser¹⁹. We sampled the QPD at 4 kHz, the smallest detectable bead displacement was ~5 nm, the smallest force measured was ~0.03 pN, and the free bead relaxation time was <20 ms. At the end of each experiment, the QPD and trap stiffness were calibrated with the same bead by using an acousto-optical deflector (IntraAction) and the equipartition method¹⁹. Programs for calibration and instrument control were written in LabVIEW 6i (National Instruments). Specimen temperature was regulated to $32.0 \pm 0.5^\circ\text{C}$ by home-made stage and objective lens (Biotech) heaters. Samples were maintained in custom-built chambers made from microscope slides that were etched to a depth of ~40 μm over an area just smaller than the coverslip, which was affixed by double-stick tape. The volume of the resulting chamber was 15–20 μl . Solution exchange was driven by a PicoPlus pump (Harvard Apparatus) with thin polyethylene tubes and a valve controller (Warner Instruments).

Reagents and experimental conditions. Tubulin was purified from cow brain by thermal cycling and chromatography, and then labelled with rhodamine or biotin²⁶. Pellicles were prepared from *Tetrahymena* and affixed to glass coverslips⁶. Glass microspheres (Polysciences) were coated further with streptavidin conjugated to bovine serum albumin²⁶ or with MT-associated proteins prepared by boiling the brain proteins that eluted in high salt buffer from the phosphocellulose column used for tubulin purification. Labile MTs were grown from biotinylated and native tubulin (ratio 1:5 to 1:10, total concentration ~1.3 mg ml⁻¹) in 80 mM PIPES buffer (pH 6.9), 1 mM EGTA, 1–4 mM MgCl₂ and 2 mM dithiothreitol, 0.2–0.5 mg ml⁻¹ of casein and 1 mM GTP. MT growth from pellicles was confirmed by differential interference contrast imaging with a CCD Cascade camera (Roper Scientific). When the MTs were ~15–20 μm , we washed in (at a rate of 30 $\mu\text{l min}^{-1}$) about three chamber volumes of a mixture of rhodamine-labelled and native tubulin (ratio 1:3, total concentration 0.4 mg ml⁻¹) with 1 mM GMPCPP (Jena Bioscience); this formed MT 'caps' of 1–4 μm . Immediately thereafter, the chamber was washed for 10 min with ten volumes of isothermal buffer containing neither tubulin nor nucleotides. In some experiments, we grew long biotinylated MTs, slowly washed in three chamber volumes of a 1.4 mg ml⁻¹ mixture of 1:3 rhodamine-labelled and native tubulin in 1 mM GTP without biotin, and then immediately capped the MTs with unlabelled GMPCPP tubulin. This created MTs that looked in Rhodamine channel like those made by the first method, but which fell apart noticeably faster on illumination. Similar results were obtained with both methods, indicating that the exact position and composition of the photosensitive MT segment were unimportant.

Beads were introduced by flow. The chamber was then inverted for 2–3 min to facilitate binding of the beads to biotinylated MT segments. The chamber was returned to its upright orientation and washed briefly. Beads that remained but failed to associate with MTs settled to the bottom of the chamber. Experiments that used beads covered with MT-associated protein were identical, except that no biotinylated tubulin was used. For experiments in different Mg²⁺ concentrations we grew MTs as above, except that *Chlamydomonas* axonemes (a gift from M. Porter, University of Minnesota) were used for nucleation. After the chamber had been washed with the buffer described above, we exchanged it with a buffer that was identical except for different concentrations of MgCl₂. Figure 3c shows result from 16–19 force measurements for each indicated Mg²⁺ concentration (a total of 30–43 observations in each group). Measurements at 1 and 4 mM Mg²⁺ produced statistically indistinguishable results and thus were grouped.

Depolymerization of MTs was induced with the microscope's epi-illumination system, which directed light that stimulated rhodamine fluorescence onto

one portion of the pellicle under observation. The average distance from a trapped bead to the edge of the pellicle was $3.8 \pm 0.2 \mu\text{m}$, which is significantly less than a typical MT; thus, our observations are unlikely to have been affected by the transient presence of the rhodamine-labelled segment. Moreover, in several experiments the MT-associated bead detached before the green light was turned on (as the GMPCPP cap is not perfectly stable). Results from these events were not detectably different, suggesting that the illumination regimen did not create artefacts.

Criteria for choosing beads for study. Normally the trapped bead showed brownian movements, the amplitude of which depended on the stiffness of the trap (Fig. 2a). In nine experiments selected at random, the average trap stiffness was 0.0083 pN nm⁻¹ in the plane perpendicular to the optic axis; the average amplitude of brownian motion for a bead in this trap was $23.2 \pm 2.1 \text{ nm}$ (for a frequency range of 0.1–100 s⁻¹). Binding to MTs longer than ~2 μm did not alter these values detectably. To simplify data analysis we selected beads attached to MTs that were oriented roughly along one of the QPD axes. We routinely tested the attachments by moving the piezo nanopositioning stage (Physic Instrumente) in 0.05- μm steps both parallel and perpendicular to the MT axis. A properly attached bead was displaced from the trap's centre only with movements parallel to the MT. Such tension decreased the amplitude of the bead's movement, but only in the direction parallel to the MT ($55 \pm 12\%$ per 1 pN of tension; Fig. 2b). Beads that behaved differently from the above were not included in case they were attached to several MTs with different orientations.

Received 5 May; accepted 8 August 2005.

- Inoue, S. & Salmon, E. D. Force generation by microtubule assembly/disassembly in mitosis and related movements. *Mol. Biol. Cell* **6**, 1619–1640 (1995).
- Dogterom, M., Kerssemakers, J. W., Romet-Lemonne, G. & Janson, M. E. Force generation by dynamic microtubules. *Curr. Opin. Cell Biol.* **17**, 67–74 (2005).
- Waterman-Storer, C. M. & Salmon, E. D. Endoplasmic reticulum membrane tubules are distributed by microtubules in living cells using three distinct mechanisms. *Curr. Biol.* **8**, 798–806 (1998).
- Waterman-Storer, C. M., Worthylyke, R. A., Liu, B. P., Burridge, K. & Salmon, E. D. Microtubule growth activates Rac1 to promote lamellipodial protrusion in fibroblasts. *Nature Cell Biol.* **1**, 45–50 (1999).
- Koshland, D. E., Mitchison, T. J. & Kirschner, M. W. Polewards chromosome movement driven by microtubule depolymerization *in vitro*. *Nature* **331**, 499–504 (1988).
- Coue, M., Lombillo, V. A. & McIntosh, J. R. Microtubule depolymerization promotes particle and chromosome movement *in vitro*. *J. Cell Biol.* **112**, 1165–1175 (1991).
- Lombillo, V. A., Nislow, C., Yen, T. J., Gelfand, V. I. & McIntosh, J. R. Antibodies to the kinesin motor domain and CENP-E inhibit microtubule depolymerization-dependent motion of chromosomes *in vitro*. *J. Cell Biol.* **128**, 107–115 (1995).
- Lombillo, V. A., Stewart, R. J. & McIntosh, J. R. Minus-end-directed motion of kinesin-coated microspheres driven by microtubule depolymerization. *Nature* **373**, 161–164 (1995).
- Molodtsov, M. I. *et al.* A molecular-mechanical model of the microtubule. *Biophys. J.* **88**, 3167–3179 (2005).
- Molodtsov, M. I., Grishchuk, E. L., Efremov, A. K., McIntosh, J. R. & Ataullakhanov, F. I. Force production by depolymerizing microtubules: a theoretical study. *Proc. Natl Acad. Sci. USA* **102**, 4353–4358 (2005).
- Muller-Reichert, T., Chretien, D., Severin, F. & Hyman, A. A. Structural changes at microtubule ends accompanying GTP hydrolysis: information from a slowly hydrolyzable analogue of GTP, guanylyl (α,β)methylene diphosphonate. *Proc. Natl Acad. Sci. USA* **95**, 3661–3666 (1998).
- Gigant, B. *et al.* The 4 Å X-ray structure of a tubulin:stathmin-like domain complex. *Cell* **102**, 809–816 (2000).
- Caplow, M., Ruhlen, R. L. & Shanks, J. The free energy for hydrolysis of a microtubule-bound nucleotide triphosphate is near zero: all of the free energy for hydrolysis is stored in the microtubule lattice. *J. Cell Biol.* **127**, 779–788 (1994).
- Howard, J. & Hyman, A. A. Dynamics and mechanics of the microtubule plus end. *Nature* **422**, 753–758 (2003).
- Mandelkow, E. M., Mandelkow, E. & Milligan, R. A. Microtubule dynamics and microtubule caps: a time-resolved cryo-electron microscopy study. *J. Cell Biol.* **114**, 977–991 (1991).
- Hyman, A. A., Salsler, S., Drechsel, D. N., Unwin, N. & Mitchison, T. J. Role of GTP hydrolysis in microtubule dynamics: information from a slowly hydrolyzable analogue, GMPCPP. *Mol. Biol. Cell* **3**, 1155–1167 (1992).
- Vigers, G. P., Coue, M. & McIntosh, J. R. Fluorescent microtubules break up under illumination. *J. Cell Biol.* **107**, 1011–1024 (1988).
- Allersma, M. W., Gittes, F., deCastro, M. J., Stewart, R. J. & Schmidt, C. F. Two-dimensional tracking of ncd motility by back focal plane interferometry. *Biophys. J.* **74**, 1074–1085 (1998).
- Visscher, K. & Block, S. M. Versatile optical traps with feedback control. *Methods Enzymol.* **298**, 460–489 (1998).

20. Svoboda, K. & Block, S. M. Force and velocity measured for single kinesin molecules. *Cell* **77**, 773–784 (1994).
21. Westermann, S. *et al.* Formation of a dynamic kinetochore-microtubule interface through assembly of the Dam1 ring complex. *Mol. Cell* **17**, 277–290 (2005).
22. Gildersleeve, R. F., Cross, A. R., Cullen, K. E., Fagen, A. P. & Williams, R. C. Jr Microtubules grow and shorten at intrinsically variable rates. *J. Biol. Chem.* **267**, 7995–8006 (1992).
23. O'Brien, E. T., Salmon, E. D., Walker, R. A. & Erickson, H. P. Effects of magnesium on the dynamic instability of individual microtubules. *Biochemistry* **29**, 6648–6656 (1990).
24. Walker, R. A. *et al.* Dynamic instability of individual microtubules analyzed by video light microscopy: rate constants and transition frequencies. *J. Cell Biol.* **107**, 1437–1448 (1988).
25. Fygenson, D. K., Braun, E. & Libchaber, A. Phase diagram of microtubules. *Phys. Rev. E* **50**, 1579–1588 (1994).
26. Hunt, A. J. & McIntosh, J. R. The dynamic behaviour of individual microtubules associated with chromosomes *in vitro*. *Mol. Biol. Cell* **9**, 249–261 (1998).
27. Miranda, J. L., De Wulf, P., Sorger, P. & Harrison, S. C. The yeast DASH complex forms closed rings on microtubules. *Nature Struct. Mol. Biol.* **12**, 138–143 (2005).
28. Nicklas, R. B. How cells get the right chromosomes. *Science* **275**, 632–637 (1997).
29. McIntosh, J. R., Grishchuk, E. L. & West, R. R. Chromosome–microtubule interactions during mitosis. *Annu. Rev. Cell. Dev. Biol.* **18**, 193–219 (2002).
30. Brouhard, G. J., Schek, H. T. III & Hunt, A. J. Advanced optical tweezers for the study of cellular and molecular biomechanics. *IEEE Trans. Biomed. Eng.* **50**, 121–125 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Hunt and G. J. Bouchard for sharing the plans for their laser tweezers; T. Perkins for advice, the quadrant photo detector design and programs to help with instrument calibration; H. Higuchi for tips on buffers; and members of McIntosh laboratory, and A. I. Vorobjev and G. P. Georgiev for help and support. V. Sarbash, T. Buxkemper and C. Bowen helped with building the trap. This work was supported in part by grants from the NIH to J.R.M., who is a Research Professor of the American Cancer Society.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.M. (richard.mcintosh@colorado.edu).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Where East meets West

After several years of decline, the number of foreign students enrolling in US graduate schools rose by 1% this autumn, according to the latest report from the Council of Graduate Schools. But it is not yet clear whether this represents a true reversal in fortunes.

The influx of foreign students is an important indicator because it offers an insight into the scientific workforce of the future. In the United States, that workforce has become increasingly international as the number of US students pursuing graduate studies has steadily declined — especially in the hard sciences. This fall was readily balanced out by foreign students, particularly from China, coming to the country to study. But after visa restrictions were tightened in the wake of the terrorist attacks of 2001, there were signs that foreign students were looking elsewhere: Chinese applications for study at US graduate schools fell by 45% between 2003 and 2004.

Of course, visa restrictions are only part of the story. China, Japan, Singapore and South Korea have all cranked up the competition in recent years. China is building research institutions at its major universities

and is undertaking a recruiting drive to match. Korea is using its freedom from US stem-cell politics to attract interest and establish itself at the forefront of that technology. Japan is strengthening efforts to help postdocs find jobs on its soil. And Singapore is drawing on international collaborations to attract attention, and has recently hosted the first Keystone conference to be held outside North America.

Such increased competition is a positive thing. It will force Western universities to keep their programmes vital and lead them to look for more ways to make foreign scientists feel welcome, rather than taking them for granted. And the changes in the East offer fresh alternatives to scientists who once saw the West as their only option for study and work.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

European Head Office, London
 The Macmillan Building, 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director:
 Nevin Bayoumi (4978)
European Sales Manager:
 Andy Douglas (4975)

Natureevents: Sille Opstrup (4994)
UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
Germany/Austria/The Netherlands:
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
Germany/Austria/
The Netherlands:
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 345 Park Avenue South,
 10th Floor, New York,

NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Balancing act

Mounting responsibilities can swamp the newly independent scientist. **Kendall Powell** asks if it's possible to manage your time without losing your creativity.

Christina Hull rarely drops the ball. After two years as an assistant professor at the University of Wisconsin, Madison, she has her lab up and running, teaches undergraduate courses, serves on several committees and still manages to find time for the gym and for dinner every night with her six-year-old daughter, Madeline, and her husband, industry scientist Rob Brazas. But she's put one of her favourite leisure activities — playing the piano — on hold.

How does she do it all? She and her husband stick to a schedule that revolves around Madeline's school and daycare. That keeps everyone on track to be home for dinner and ensures that Hull works efficiently during the day. But, she admits, it isn't a perfect system.

"I did no time management when I was teaching a new course. I got in no workouts, cooked no dinners, and only slept from about 2 a.m. to 6 a.m. for those six weeks." Luckily, her family's schedule does allow for flexibility when one spouse needs it.

But as the responsibilities of being a new investigator pile up, forsaking sleep will only get you so far. Time management is a skill all academics have to master early in their careers. By year two, a new lab should be equipped and staffed to run smoothly on a daily basis, freeing up time to refocus mental energy on research goals and bench work.

But by year two, non-research responsibilities such as committee work and teaching have also ramped up. With only 24 hours in a day, it can be hard to carve out time to help lab workers, write grants and manuscripts, and just think about projects amid all the pressing but tenure-neutral matters you're also dealing with. No matter how well you serve on committees or advise undergraduates, by about year seven your career will depend on publishing and advancing deep science. Securing tenure takes an enormous investment of wisely managed hours early on.

Watching the clock

Administrative duties such as interviewing and grant-writing dominate the first year or so for most new investigators. But as the lab settles in, they find long stretches of time harder to come by — just as they are ramping up their experiments.

Joaquin Espinosa, a molecular cell biologist starting his second year at the University of Colorado, Boulder, says it can be difficult to switch from managing time in monthly blocks to managing it by the day or hour. "I'd better start learning how to do these big things in small chunks of time," he says.

Hull has made that transition. Every morning, she asks herself: "What do I have to do today to make my lab run?" and starts organizing her day while still in the shower or commuting to work.

She and her husband each keep a detailed calendar

Both Carol Thornber (top) and Sandra Schmid divide their working weeks by categorizing their tasks.



IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

on their computers of all work and home commitments at least a few months ahead. Carol Thornber, a marine ecologist at the University of Rhode Island in Kingston, goes one step further, colour-coding her calendar into research, teaching, other work and home categories. This gives her an immediate picture of whether she has balanced her week and also tells her when she still has free time for writing or preparing lectures.

Thornber asks her colleagues to bug her when manuscripts slip down her to-do list, so she can make them a priority. Both Thornber and Hull rely on their faculty mentors to guide them in weighing their campus and departmental duties against their research goals.

"Mine sat me down and while shaking his head 'No' told me: 'You need to practice this motion,'" says Thornber. Hull sought guidance for how many papers she should peer-review in a semester, when to turn down reviews and how much time to spend on committee work.

"You have to be a little bit selfish at this stage," she notes. "Do the maths — how long does it take you to review a paper? Am I using my time during the week to move my lab forward?"

Sandra Schmid, a cell biologist at the Scripps Research Institute in La Jolla, California, suggests using a time-management matrix. Tasks are divided into two columns, important or not important, and two rows, urgent and not urgent.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

"If you are focused on the tenure clock, that's what goes in the important, not urgent box — getting three papers published, giving seminars, getting funded," she says. "Focus towards accomplishing those things and then go home at six."

Inevitably, a young lab head will get caught up in immediate, seemingly urgent tasks such as purchasing or paperwork. Junior faculty members agree that it's only natural to attend to such matters when setting up — but they warn the new investigator to maintain focus on longer-term goals, or risk becoming a great lab manager and a poor planner of original research.

Not enough hours

Daily admin tasks, service work and teaching can eat up time. Paperwork and committee meetings may not be glamorous, but they come with the territory. To keep a lid on these potential time drains, you need to work out short cuts and self-imposed rules.

Bernard Golding, an organic chemist at the University of Newcastle upon Tyne, UK, says handle paperwork promptly to avoid "worrying about the nasty tasks you haven't done". He also urges committee members to rethink the habit of scheduling meetings in hour-long units. Can a quick agenda be covered in 15 minutes? Can a decision be trusted to a select few? Can a consensus be reached by e-mail?

Schmid approaches service opportunities as ways to advance her influence in both her department and

Sam Sia (top) has tamed his lectures and Dirk Schübeler watches his biological clock to keep his research on track.



field. She recalls being asked to serve on the editorial board of a respected journal after about four years as an assistant professor. She agreed — but immediately stopped reviewing papers for other journals. She saw the decision as 'trading up' in responsibilities. She also found that being a "fully engaged, participating member" of one committee helped her to decline invitations to join others without suffering any political fallout.

Likewise, junior faculty members should dedicate significant energy to their teaching responsibilities. After all, Golding points out, high-quality teaching can attract the best students to your group and can influence the direction of your research. Although preparing lectures for a first course can be daunting, it should not take inordinate amounts of time.

Sam Sia, a bioengineer at Columbia University in New York, found a simple way to ensure that his undergraduate thermodynamics course would not take over his first semester. "I use the chalkboard. No handouts and no fancy slide-show presentation," he says. He also advises "pick a good textbook and stick to it", using its standard problems and diagrams. Students, he notes, appreciate clear, organized lectures and do not need originality.

Creative juices

As Espinosa points out, once the lab is functioning, planning experiments and pondering projects gets squeezed into shorter, more frequent time slots. New investigators find this adjustment particularly tricky. Tap into when you do your best thinking and then talk to lab members and do other daily jobs outside those sacred times.

Dirk Schübeler, an epigenetics expert at the Friedrich Miescher Institute in Basel, Switzerland, says the hardest part of the pre-tenure position is maintaining creativity under time pressure. Coming out of a meeting and switching into creative mode doesn't come naturally to most people, he says, so you have to reserve time to get into the groove.

"Follow your biological clock and recognize which parts of the day you are more creative than others," he says. "Then keep those hours open for thinking." Thornber and Golding reserve one day a week or month to stay home and work on creative tasks.

As he grew busier, Golding began scheduling formal meetings with his trainees each week. Schmid urges investigators to keep lab workers informed of grant aims and timelines. Lab workers should be trained to be self-sufficient and goal-oriented, Schmid says. "Don't agonize about a person who is not producing," she warns. Instead, ask to see improvement in the next three months.

Schmid and Hull say that not all the ingredients of a successful career need to happen simultaneously in the first five years. But planning and organization help ensure they will come to fruition eventually.

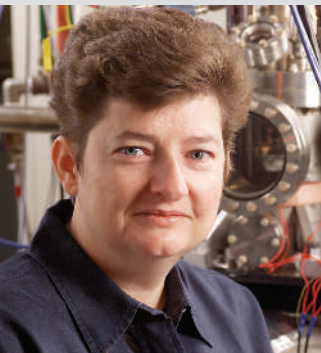
As for leisure activities, some use the prospects of engaging in them again as a way to stay motivated in the lab. Hull expects to take up the piano once she catches up.

"I've got my eye on a baby grand that in a few years from now will make a great tenure present for myself," she says.

Kendall Powell is a freelance science writer based in Broomfield, Colorado.

MOVERS

Melissa Hines, director, Cornell Center for Materials Research, Ithaca, New York



1994–present: Assistant professor, associate professor and, since 2004, professor, Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York
1992–94: Postdoctoral member of technical staff, Optical Physics, AT&T Bell Laboratories, Murray Hill, New Jersey

Melissa Hines has no idea what set her on a career path towards materials science. But her desire to be a chemist has not wavered since she was seven years old. This drive has helped to propel her to leadership at an interdisciplinary research centre comprising some 100 faculty members from Cornell University in Ithaca, New York.

As a chemistry undergraduate at the Massachusetts Institute of Technology (MIT) in Cambridge, she followed prudent advice: work for a young faculty member who will have more time to devote to his or her students. Her choice was Sylvia Ceyer, who was not only the youngest member of MIT's chemistry department at the time, but one of the few women. Ceyer inspired Hines to work hard and to focus on surface chemistry. Ceyer also highlighted the importance of communicating science to the public.

"She convinced me that if you can't explain what you are doing to someone outside the field, you probably don't understand it yourself," says Hines.

After a PhD in chemistry at Stanford University, Hines went on to a postdoc at Bell Labs in Murray Hill, New Jersey — a career-altering experience. There, she learned how materials science could be used to solve problems in fields that she never would have encountered otherwise, such as the engineering and design of microprocessors.

She has combined that appreciation for applied research with her zeal for basic science in her current position as director of the Center for Materials Research at Cornell, which emphasizes collaborations between ten different departments at the university.

In this job, Hines finds herself revisiting the advice she got at MIT about communicating with the public. "People think of chemicals as something bad, to be avoided, and that's a real challenge that the scientific community needs to address," she says. To meet that challenge, she gives seminars on public speaking for scientists to help them communicate with students, politicians and others.

Hines counsels students to keep an open mind about what kind of scientist they are and want to be. It seems odd advice from someone with such a focused career path, but she says that she has become much more open-minded about different research fields over the years.

"Life throws you a lot of opportunities and if you automatically shut these out, you could miss something wonderful," she says.

Virginia Gewin

RECRUITERS & ACADEMIA

Lessons in professorship

Scientists spend a lot of time in the university system, but few know what it means to actually work in academia as what, in North America, is described as a 'professor'. Where do you learn how to be a professor? What exactly is tenure track and how does one 'get on it'? How do you negotiate for what you need?

FORWARD to Professorship is a workshop spread over two and a half days that helps faculty-to-be and new tenure-track faculty members explore the above questions and much more. About a quarter of the workshop is devoted to issues specific to women and other under-represented groups in science, engineering and mathematics. We aim to equip participants with a personal plan for their path to tenure and to a balanced life. With funding from the US National Science Foundation, we have held the workshop for the past three years at Gallaudet University in Washington DC, with 40 to 50 participants from across the United States. We took it on the road for the first time this year, to the Massachusetts Institute of Technology.

We address the three components of professorship: teaching, research and service. For teaching, we demonstrate a variety of styles. For research, the focus is on securing funding; guest speakers from various funding agencies discuss requirements and expectations for

proposals. For service, we address the need for university service, while cautioning against over-commitment.

We also give participants time to do some writing — typically on research or teaching statements — and to receive feedback. Experts in negotiation discuss why women often do not negotiate well or much, and review strategies. Family issues such as your options if you want children and if your partner is in the same field are addressed by professors who have successfully handled these situations. Participants sit with department chairs and deans in small groups to ask questions about how administrators view certain situations.

Follow-up surveys have shown that networks and mentoring begun at the workshop continue long after it ends. One participant arrived at the workshop three years ago having decided to quit her PhD programme. While there, she found mentors and support. We are pleased to report that she has now completed her doctorate.

For more information about attending the workshop or bringing it to your institution, please contact us at Forward.office@gallaudet.edu.

Charlene Sorensen is a professor at Gallaudet University in Washington DC and a co-organizer of the FORWARD to Professorship workshop.
 ♦ www.seas.gwu.edu/~forward/advance

GRADUATE JOURNAL

Learning from teaching

I used to teach ballroom dancing as a way to earn some pocket money during my master's studies. It was an interesting experience, but I never thought of it as serious professional training, because it was so different from my real job as a scientist. But now that experience has become useful to me in the lab.

I never thought there would be a day when I would be teaching students at my university. The prospect seemed distant and unreal. But now it is a task I must undertake: I am helping to supervise a master's student. This is not a classroom full of children, but one person sitting next to me.

I suddenly realized that I bear a great deal of responsibility for the quality of his work, and that I must do my best to encourage him and to help him improve his skills. Although this scared me at first, I began to teach and to learn from teaching. I finally understood the importance of the management, psychology and philosophy classes that I had to take earlier in my education. These classes showed me a different way of looking at the world — and at people.

I now see that my experience as a dance teacher was not a waste of time either. It showed me how different people react to different situations. I've been able to transfer the knowledge gained in dance classes to laboratory teaching, and this has also helped my everyday relations with people in general.

Karolina Tkaczuk is a graduate student at the Technical University of Lodz, Poland.

Perchance to dream

Out of sight, out of mind.

Robert A. Metzger

There is a place to trade, on the third level, antique hardware in exchange for real food. But first I must walk this stretch of the seventh level in order to reach the Up-Tube. The haze is thick, inquiry motes swirling about me. I let them ask their questions but, as if in response, I inform them that they sample nothing but air. I study everything from long wavelength radio to short ultraviolet spikes, sniffing the pheromone bouquet and sampling the organic debris. I do not transmit, do not even reflect. I am operating in full stealth, invisible to the Dreamers about me, those inhabiting virtual worlds. Shops adorned with wrought iron and blossoming wisteria line the street, screaming with a full spectrum onslaught, begging me to enter and sample their virtual goods. Above hangs a golden sun. The street is undoubtedly thick with people, but I choose not to experience them, my stealth suit not only rendering me invisible, but also guiding me around people without the need of conscious intervention.

I am not a Dreamer. I am awake, safe in my stealth suit, isolated from the fantasy that has consumed the world.

"You in full stealth again, Grandpa?"

I focus, narrowing input to the merely visible, nulling the informational torrent, blossoming wisteria and wrought iron fading away as I drop the suit's human filter. In front of me stands a boy, naked, dirty, brown hair hanging in tangles. His eyes sparkle like diamonds, the Ocs wedged behind his corneas both spewing and gulping data — his gateway to Dream-generated worlds. He stands before me on the narrow sidewalk, naked people shuffling by us like water flowing around two rocks in a stream. The shops along this stretch of tunnel are choked with bodies, and nothing else — the only goods available on this level virtual — photons being the most cost-effective consumer objects. I look up. There is no sun, just a warren of steel strut and old plumbing hanging from cracked concrete. Flickering fluorescents cast everything in harsh light. We are seven levels below the surface,

in the lowest rungs of a Dreaming World, where the resource-challenged can only afford photon-based goods.

"Stealth suits are 20 years extinct, Grandpa," he says, waving a hand at me. "Isolation, anonymity, individualism, all such sad Dreams."

My suit might be 20 years old, but it gets the job done. I run a diagnostic. I am emitting nothing but a blast of infrared through my rear radiator fins, my ultraviolet ionizers at exhaust ports shedding any leaking DNA, and biometric compensators continually randomizing my movements. The motes of inquiring dust that choke this tunnel, transmitting torrents of data between them, nibble at my suit, questioning, probing, but my suit informs them that there is nothing there.

I am not like this boy, like any others on the seventh level. I am not a consumer of photons and Dreams. I am the last from a world now gone. I am flesh and blood, and I am invisible.

And yet this boy can somehow see me. How?

My suit accesses the boy, allowing me to see through his Ocs. There is no such thing as privacy in the Dreaming World — privacy inhibits the flow of commerce and Dreams. Perspective slips as I enter into the boy, his hands are now mine, adorned with jewels, gold rings, skin plastered with morphing displays, sleeves of silk covering my arms. This is the world he experiences. About me swirls frenzied motion and colours, bustling bodies loaded down with packages, bright sun above, shops adorned with wrought iron beckoning, and everywhere blossoming wisteria.

In front of me stands an old man — nearly naked, cloth tied around his waist, a white beard, aged yellow, hanging midway down his chest, his skin wrinkled, nearly translucent, thick blue veins visible across his bald head.

"I'm in full stealth," says the old man standing in front of me. "I'm not a part of your Dreaming World. I'm hidden and safe." He grins. He has no teeth. His eyes sparkle like diamonds, the Ocs wedged behind his corneas glistening. He is Dreaming of stealth suits. I reach for the old man, taking one of his gnarled hands. "You shouldn't be out here, all alone Grandpa," I hear myself say.

Then I blink and am back in my own skull, the boy holding my hand. "My stealth suit?" I ask, not understanding how it could be gone, as I look down at my nearly naked body. "I was going to make a trade on the third level for food," I say.

"There is no third level, Grandpa," he says, and sweeps a hand in front of me. "Only here. Everything else is a Dream."

I remember, knowing what he says is true, the confusion lifting a bit.

"Time to get you home, Grandpa," he says.

I nod, hoping that I will find my stealth suit there. We walk down the street, the warm sun shining down on us, the shops beckoning, full of wonderful Dreams for sale. The wisteria is in full bloom. ■

Robert A. Metzger is a hard-SF writer and a research scientist in the area of semiconductor thin films from North Carolina. His latest novel, released by Ace in 2005, is *Cusp*.



JACEY